Capítol 3. L'obtenció de recursos textuals

1. La recuperació d'informació: l'obtenció de dades al *World Wide Web*

Beynon-Davies (1993) estableix una diferenciació entre dades, informació i coneixement, conceptes que en l'àmbit de la documentació estan íntimament relacionats. Neumüller (2001) resumeix la diferenciació establerta per Beynon-Davies de la manera següent:

> "— Data is facts. A datum, a unit of data, is one or more symbols that are used to represent something.
>
> — Information is interpreted data. Information is data placed within a meaningful context. [...]
>
> — Information is necessarily subjective. Information must always be set in the context of its recipient. The same data may be interpreted differently by different people depending on their existing knowledge.
>
> — Knowledge is derived from information by integrating information with existing knowledge." (Neumüller, 2001)

Segons aquesta gradació, doncs, el traductor, en documentar-se per a una traducció tècnica o científica, sovint cerca informació que li permeti adquirir coneixements factuals en l'àmbit en què s'emmarca la traducció que ha de dur a terme. Pocs cops cerca dades segons les entén Beynon-Davies, ja que normalment s'ha d'atenir a les dades del text original que ha de traduir, tot i que sempre hi ha excepcions. Generalment, el traductor cerca informació que li permetrà entendre el missatge original i que, per tant, li suposarà un aportament de coneixements relatius a la matèria, un aportament cognitiu, mentre que, alhora, adquirirà tota una sèrie de coneixements lingüístics que li permetran elaborar un text adequat tenint en compte els

condicionats i les característiques de la llengua i la cultura d'arribada, així com de l'encàrrec de traducció.

Sovint el procés és l'invers, és a dir, mentre el traductor ha de procurar adquirir els recursos lingüístics que necessita per fer una traducció, també adquireix coneixements sobre la matèria.[1] (En el capítol 5 "El text paral·lel en la traducció especialitzada" s'aprofundeix en les necessitats i les estratègies de cerca d'informació del traductor.)

En aquest apartat, dedicat a la recuperació d'informació, es presenta en primer lloc una valoració dels recursos textuals que es poden recuperar en l'entorn Web, basada en estudis i criteris d'avaluació de recursos digitals, per tal de caracteritzar l'objecte d'estudi. A continuació, i després de comentar els principis fonamentals de recuperació d'informació per tal de poder-la recuperar posteriorment, se centra en la cerca i recuperació d'informació textual en el *World Wide Web* i en els diferents mecanismes que s'hi poden utilitzar.

1.1. Avaluació de la informació al *World Wide Web*

En el seu article "Information Literacy: The Web is not an Encyclopaedia", Larsen introdueix el concepte d'alfabetització informativa (*information literacy*), que defineix com la capacitat de saber cercar la informació que es necessita, trobar-la, avaluar-la, processar-la i utilitzar-la per prendre decisions adequades en la vida de cadascú, tot referint-se a l'entorn web com a mitjà d'informació en què aquest procés té lloc (Larsen, 2001); precisament els traductors han d'enfrontar-se amb aquesta tasca diàriament al llarg de la seva vida professional.

[1] De fet, la informació sobre qualsevol matèria té forma lingüística, per la qual cosa es fa difícil separar la informació lingüística de la factual, ja que, si més no en els textos, és la forma lingüística la que vehicula la informació que finalment, i un cop interpretat, es convertirà en coneixement.

La correcta utilització de la informació publicada al *World Wide Web* obliga a conèixer-ne la naturalesa, per tal de no esperar més, ni menys, d'allò que realment ofereix. El *World Wide Web* es coneix sobretot per ser:

- Un mitjà d'abast global i, per tant, immens.
- No estar sotmès al control de cap organisme.
- Poder ser utilitzat per transmetre informació en qualsevol de les seves morfologies.
- Accedir-hi fàcilment, si hom té l'equipament necessari.
- Oferir una gran varietat de serveis. (Chowdhury, 1999: 395)

La publicació a Internet[2] ha democratitzat la propietat, la distribució i la recuperació d'informació. Seguint amb l'argumentació de Larsen, això no obstant, no vol dir que el WWW sigui una gran enciclopèdia (entesa com conjunt organitzat d'articles escrits i revisats per experts), ni tan sols una gran biblioteca (entesa com a conjunt organitzat d'obres independents publicades seguint un procés editorial). El web és, en realitat, un dipòsit en format digital de llibres, conjunts de dades, enciclopèdies, biblioteques així com "any disparate piece of text, graphic, or sound byte that someone chose to put on-line" (Larsen, 2001). Això és el que fa que molts internautes pensin que sempre disposen de la informació a la punta dels dits, fet que Codina ha batejat de *ciberingenuïtat* (1995), atès que sembla un raonament no gaire justificat en la pràctica, puix que no sempre es troba la informació que es necessita.

Per tal de saber, doncs, què cap esperar del *World Wide Web* cal conèixer com és: quanta informació s'hi pot trobar, com s'hi distribueix i quins tipus de recursos inclou, aspectes que s'abordaran a continuació.

[2] El terme Internet sovint s'utilitza com a sinònim de *World Wide Web*, com en aquest cas, però en realitat el WWW és un subconjunt d'Internet que es regeix pel protocol HTTP. A Internet també es troben altes components amb protocols diferents, com l'FTP, el correu electrònic, els grups de notícies, la connexió Telnet i el Gopher.

1.1.1. Distribució de la informació a la World Wide Web

El web de pàgines públiques i de fàcil accés, que és la seva part més coneguda (en contraposició, vegeu el punt c) "El web invisible" de l'apartat 1.1.3 "Els cercadors segons el seu funcionament", pàgina 156), estava format, segons l'estudi *Sizing the Internet* realitzat per Cyveillance el juliol de 2000 (Cyveillance, 2000) per aproximadament 2.100 milions de pàgines, i augmentava a raó de 7 milions de pàgines per dia, per la qual cosa s'esperava que dupliqués el seu volum a començaments de 2001, superant els 4.000 milions. Les pàgines són arxius que, com a mitjana, tenen un volum de 10 quilobytes, i contenen uns 23 enllaços interns, 6 d'externs i 14 imatges. Estudis duts a terme per altres entitats arriben a conclusions similars a les d'aquest estudi.

La distribució de totes aquestes pàgines és desigual des de molts punts de vista. Per exemple, l'esmentat estudi de Cyveillance afirma que el 84% de les pàgines del *World Wide Web* resideixen en servidors dels Estats Units, la qual cosa palesa un gran desequilibri geogràfic en l'origen dels recursos.

Des d'un punt de vista estructural, aquest ingent volum d'informació, organitzat de manera jeràrquica en pàgines que formen part de documents, que alhora es troben en servidors, no guarden cap mena d'ordre sistemàtic; tanmateix, sí que semblen comportar-se en l'entorn web seguint uns determinats patrons. Això almenys és el que reflecteix l'estudi patrocinat per les empreses AltaVista, Compaq i IBM (Broder *et al.*, 2000) sobre la topografia del Web. En aquest estudi es resumeix l'estructura del *World Wide Web* en tres subconjunts vinculats entre ells formant un corbatí:

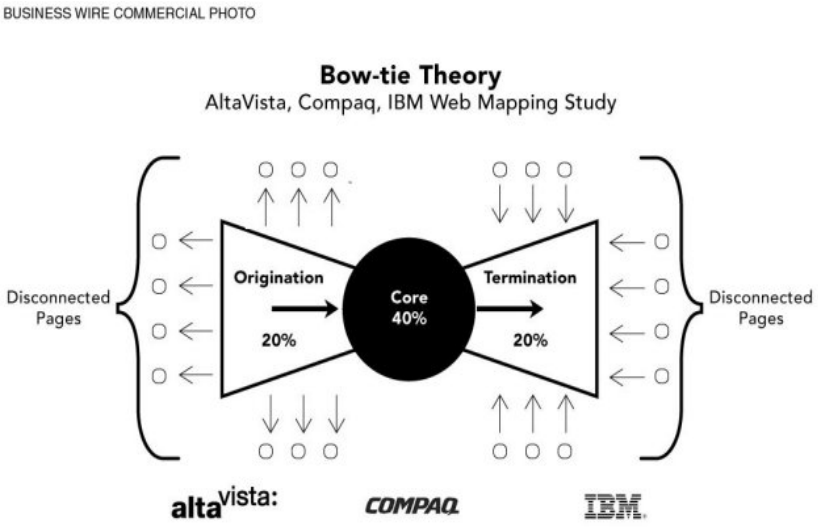

Figura 3-1. Esquema de la topografia de la *World Wide Web* segons (Broder *et al.*, 2000).

El primer dels subconjunts, el nus del corbatí, està format aproximadament pel 40% dels documents de l'entorn web, altament interrelacionats entre ells, és a dir, que permeten la navegació entre diferents pàgines i tornar a la pàgina d'origen[3]. El subconjunt que conforma el llaç de l'esquerre, que engloba una cinquena part del web, està format per pàgines d'origen, des de les quals es pot arribar al nus, però a les quals no es pot arribar des d'un document del nus. Habitualment són pàgines de nova creació que encara no han atret l'interès d'altres internatues, o que només reben enllaços interns. El cas del subconjunt que conforma el llaç dret, que també engloba una cinquena part del web, és precisament el contrari: són pàgines destí, que reben enllaços des de documents del nus, tot i que no permeten fer el camí enrere perquè no contenen enllaços cap a aquests últims.

El conjunt restant, un 20% del total del web, està format per pàgines desconnectades del nus, tot i que poden estar connectades amb les pàgines origen o destí. D'altra

[3] Com es veurà en el proper apartat (1.1.2 "Tipus de recursos web i qualitat de la informació", pàgina 127), el fet que una pàgina rebi molts enllaços implica en certa mesura que és un recurs de qualitat.

banda, però, segons Lawrence i Giles (1999), entre el 8 i el 10% dels enllaços que contenen les pàgines web són morts, ja sigui perquè el node destí ha deixat d'existir o perquè es va cometre un error en crear-lo. Aquestes planes alimentarien aquest 20% de pàgines poc connectades amb la resta de recursos.

Des d'un punt de vista lingüístic, la distribució de la informació no està tan equilibrada com ho està estructuralment. La llengua predominant sobre totes les altres és, sense cap mena de dubte, l'anglès, amb uns valors que no es corresponen proporcionalment amb el nombre d'internautes que tenen l'anglès com a llengua materna. En relació amb la distribució de les llengües en l'entorn web s'han realitzat diversos estudis que arriben a percentatges diferents, tot i que no gaire allunyats els uns dels altres.

En l'estudi presentat per Vilaweb a mitjan l'any 2000 (Vilaweb, 2000), el més extens que s'ha pogut consultar, s'indica una clara supremacia de l'anglès sobre totes les altres llengües:

Anglès	214.250.996	68,39%	Polonès	848.672	0,27%
Japonès	18.335.739	5,85%	Hongarès	498.625	0,16%
Alemany	18.069.744	5,77%	Català	443.301	0,14%
Xinès	12.113.803	3,87%	Turc	430.996	0,14%
Francès	9.262.663	2,96%	Grec	287.980	0,09%
Castellà	7.573.064	2,42%	Hebreu	198.030	0,06%
Rus	5.900.956	1,88%	Estonià	173.265	0,06%
Italià	4.883.497	1,56%	Romanès	141.587	0,05%
Portuguès	4.291.237	1,37%	Islandès	136.788	0,04%
Coreà	4.046.530	1,29%	Eslovè	134.454	0,04%
Holandès	3.161.844	1,01%	Àrab	127.565	0,04%
Suec	2.929.241	0,93%	Lituà	82.829	0,03%
Danès	1.374.886	0,40%	Letó	60.959	0,02%
Noruec	1.259.189	0,38%	Búlgar	51.336	0,02%
Finès	1.198.956	0,38%	Basc	36.321	0,01%
Txec	991.075	0,32%			

Taula 3-1. Resultats de l'estudi de Vilaweb ordenat per volum de pàgines per llengua (e. p.)

A partir d'aquestes dades s'observa un clar desequilibri, principalment entre les llengües majoritàries, el volum d'informació de les quals està molt per sota del percentatge de població real, i fins i tot virtual (població assídua a Internet), excepte en el cas de

l'anglès, en què el desequilibri és gairebé proporcionalment invers, i el japonès, que sí que manté un equilibri entre la població i el volum d'informació.

En un estudi realitzat l'any 2001, l'Online Computer Library Center (OCLC, 2001) calculà el percentatge de pàgines web d'accés públic en cadascuna de les llengües següents:

Anglès	73%	Italià	2%
Alemany	7%	Coreà	2%
Japonès	5%	Polonès	1%
Castellà	3%	Portuguès	1%
Xinès	3%	Rus	1%
Holandès	2%	Suec	1%

Taula 3-2. Resultats de l'estudi d'OCLC (e. p.)

L'estudi d'OCLC comptabilitza pàgines escrites totalment o parcialment en les diferents llengües: en el cas de les pàgines en més d'una llengua, aquestes es comptabilitzaven diverses vegades, una per cada llengua. Els resultats mostren un augment del volum d'informació en llengua anglesa, probablement fictici, atès que deixa de tenir en compte moltes llengües minoritàries.

La Fundación Redes y Desarrollo publica amb una freqüència gairebé anual els seus estudis *Lengua y Cultura* sobre la distribució de la llengua al *World Wide Web* (Funredes, 2002). Segons aquests estudis, que únicament recullen llengües europees, i sobretot les romàniques, el predomini de la llengua anglesa comença a cedir davant l'augment d'informació en altres llengües:

Anglès	54,00%
Alemany	5,69%
Castellà	4,61%
Francès	3,06%
Italià	2,81%
Portuguès	0,17%
Romanès	0,16%
Altres	24,96%

Taula 3-3. Estudi *Lengua y Cultura* de Funredes, 1998 (e. p.)

Anglès	49,00%
Alemany	7,06%
Castellà	5,68%
Francès	4,70%
Italià	3,19%
Portuguès	2,75%
Romanès	0,16%
Altres	27,46%

Taula 3-4. Estudi *Lengua y Cultura* de Funredes, 2002 (e. p.)

De totes les estadístiques que s'han presentat es desprèn que, en cercar uns continguts determinats, serà més probable trobar-los en anglès que en qualsevol altra llengua. També sembla reflectir-s'hi una certa tendència que aquesta situació està canviant, atès que la diferència entre el volum d'informació en anglès disminueix ahora que augmenten lleugerament els de les altres llengües majoritàries. Sense voler entrar-hi en detall, aquest desfasament entre llengües es deu, entre altres, al mateix desenvolupament d'Internet, que inicialment va tenir una major repercussió als EUA, com també a la voluntat dels autors dels recursos web de fer-se entendre globalment, recorrent a l'anglès com a *lingua franca*.

Un altre aspecte que tenir en compte en utilitzar el *World Wide Web* com a font d'informació i documentació és la seva **volatilitat** (Caslon Analytics, 2002), entesa des de diferents punts de vista:

- la mitjana de vida d'una pàgina és inferior a 2 anys, i la d'un lloc web la supera lleugerament;
- aproximadament el 5% de les pàgines web estan en estat intermitent: han desaparegut però tornaran a aparèixer;
- tot i que un lloc web mantingui les mateixes URL (estigui format per les mateixes pàgines) durant 2 anys aproximadament, el seu contingut pot variar en qualsevol moment;
- quan s'edita un nou material al web, sovint se'n reemplaça l'antecedent;
- tota pàgina web està en continu procés de modificació, tant tècnicament com des del punt de vista del seu contingut[4];
- les pàgines més susceptibles de sofrir modificacions són les que ocupen nivells més profunds en l'estructura d'un document.

[4] En el seu article en què descriu un estudi dut a terme sobre la persistència de les pàgines Web, Koehler assenyala que en un període d'un any més del 99% de les pàgines que formaven part del corpus objecte de l'estudi van modificar el seu contingut amb diferent freqüència.

Respecte d'això, Baeza-Yates i Ribeiro-Net (1999) afegeixen que aproximadament el 40% de les pàgines modifiquen el seu contingut cada mes. La volatilitat, tant estructural com des del punt de vista del contingut, és probablement el detonant que provoca nivells considerables de **redundància** al web. Sovint, molts llocs web, per tal d'evitar col·lapsar el servidor que l'acull o no desaparèixer de la xarxa si aquell servidor cau, es dupliquen en altres servidors, de vegades fins i tot canviant la URL. D'altra banda, el fenomen de transclusió descrit al primer apartat d'aquest capítol (1.2.2 "Les estructrues hipertextuals", pàgina 76) afavoreix la redundància d'informació en l'entorn web. Per últim, la manca de control sobre la utilització de la informació que apareix publicada en format digital i que és d'accés públic permet copiar parcialment o totalment la informació recollida en un lloc determinat i publicar-la amb una altra URL, tot i que es tracti d'una acció il·legal. Per tot això, es calcula que aproximadament el 30% de les pàgines publicades al Web són duplicats o gairebé duplicats d'altres pàgines (Baeza-Yates i Ribeiro-Net, 1999: 368).

Tant la volatilitat com la redundància són factors que provoquen una certa fragilitat física i lògica del *World Wide Web* com a mitjà d'informació (Codina, 2002: 87), i que repercutiran sobre l'efectivitat dels diferents sistemes de recuperació d'informació en aquest entorn (vegeu l'apartat 1.1.3 "Els cercadors segons el seu funcionament", pàgina 144).

1.1.2. Tipus de recursos web i qualitat de la informació

Gairebé tots els experts en el *World Wide Web* coincideixen en el fet que el seu punt fort rau en la facilitat amb què es pot publicar en aquest mitjà i la gran quantitat de públic potencial que hi pot accedir. Amb un equip informàtic estàndard, l'accés a un servidor i pocs coneixements informàtics, qualsevol persona pot editar la seva pàgina web i fer-la accessible a tot el món. Tal com s'ha indicat en el subapartat anterior, en qüestió de segons es creen, i també es destrueixen, recursos d'accés públic en format digital. Aquesta democratització de la informació afecta tant el procés d'elaboració com de distribució i consum d'informació: acaba amb els impediments, sovint econòmics, de

la producció de documents i amb les limitacions físiques de la seva distribució, per la qual cosa arriba virtualment a qualsevol lector potencial.

Actualment ha quedat palès que aquesta democratització, que inicialment es va contemplar com el punt fort de l'hipertext i de l'entorn web com a mitjà de publicació i distribució d'informació, també ha comportat certs inconvenients, relatius sobretot a la qualitat dels recursos que s'hi publiquen. Aquesta situació està provocada principalment per la manca de paràmetres de control previs a la publicació que sí que estan previstos en el procés editorial de publicació de qualsevol obra tradicional. Des de la correcció ortogràfica fins a la revisió *per peers*, l'objectiu del procés editorial és el de garantir la qualitat de la publicació, així com l'autenticitat i el caràcter únic del seu contingut, avalat i preservat pels diferents identificadors legals de l'obra (com ara l'ISBN, l'ISSN o el seu dipòsit legal).

Aquest risc, que ha derivat en una situació real, el va posar de manifest de manera molt eloqüent T. Matthew Ciolek en el seu article "Today's WWW – Tomorrow's MMM? the specter of Multi-Media Mediocrity" (Ciolek, 1997), en el qual denunciava que cada cop és més difícil trobar informació de qualitat entre tantes pàgines publicitàries i trivials (*vanity publications*).

Igual que en format paper, l'espectre de publicacions en format digital es pot considerar una mena de continuum que abraça des de la publicació més trivial fins a la més especialitzada. La diferència entre tots dos continuums, el de les publicacions en format paper i en format digital, és que el segon inclou molta més variació, a causa per exemple, de la inclusió d'elements publicitaris en recursos que no ho són, per la qual cosa resulta molt més difícil identificar els diferents tipus de recursos a la xarxa que en altres mitjans més tradicionals, com la premsa o la televisió. A més de la publicitat, la manca de dades específiques relatives al recurs, com l'autor, la data, el responsable secundari o les fonts de referència, també en dificulta la interpretació.

D'altra banda, els diferents tipus de recursos textuals tradicionals, en format paper, sovint es diferencien de manera pragmàtica per la situació comunicativa en què intervindran; aquest aspecte resulta difícil de mantenir en un entorn virtual en el qual tothom té accés a qualsevol informació en qualsevol moment i en qualsevol situació.

Arribats a aquest punt, i tenint en compte tot el que sobre el web s'ha dit fins al moment, es podria afirmar que el *World Wide Web* és tan dinàmic com caòtic, i que inclou informació de dubtosa qualitat. Així doncs, resultarà peremptori establir una tipologia de recursos web que faciliti la interpretació de la informació que contenen, que no portin a error al lector o usuari del recurs, i que facilitin l'establiment de certs criteris per determinar-ne la qualitat a l'hora de consultar-los.

Aquesta reflexió sobre la tipologia de recursos web s'ha dut a terme principalment des del món acadèmic, en resposta a les seves pròpies necessitats informatives. Inicialment es van establir categories amb criteris heterogenis, com la proposada per Tillman (1995), que dividia els recursos en trivials, literatura grisa, publicitat/relacions públiques i multimèdia. Els recursos trivials els descrivia de la manera següent:

> "A vanity work may be a very specific document that has information of great value but it hasn't been through the peer review process intrinsic to scholarship or it hasn't been disseminated by the trade publishing industry. Heretofore, vanity and short-run speciality publishing has been possible in print and can be 'quality' in nature, although its value may not be as easy to determine without analysis." (Tillman, 1995)

Segons això, gran part del web públic és trivial, i els únics recursos que s'han de tenir en compte, des del punt de vista de la seva qualitat, són aquells que funcionen com a mirall d'una publicació tradicional.

Altres classificacions, com la proposada per Ury i MacFarland (2001), ordenen els recursos en funció principalment del seu origen:

- Pàgines comercials: inclou des de màrqueting digital fins a pàgines de promoció d'empreses o entitats comercials.
- Publicacions trivials: pàgines personals, o sense revisió editorial o altres criteris que en garanteixin la qualitat.
- Literatura grisa: pamflets d'associacions professionals, informació proporcionada per organitzacions sense ànim de lucre amb l'objectiu d'informar o educar.

- Informació acadèmica: actes de congressos, articles en revistes especialitzades digitals, etc.
- Pàgines propietàries: bases de dades per subscripció, com índexs d'articles.

Les tipologies que classifiquen en funció de l'origen del recurs, com la que s'acaba de presentar, no acostumen a cobrir tot l'espectre de recursos web, atès que, per exemple, no preveuen cap categoria per a recursos com ara un diari en línia o les pàgines web de l'Administració pública.

En una altra classificació Alexander i Tate (2001) prenen com a criteri de classificació l'objectiu de l'emissor del recurs i presenten la tipologia següent:

a) Pàgines de suport i defensa (*advocacy Web pages*): patrocinades per una organització amb l'objectiu d'influir sobre l'opinió pública, de convèncer ideològicament. Aquestes pàgines acostumen a estar sota un domini .org (organització).

b) Pàgines publicitàries i de negocis (*business/marketing Web pages*): patrocinades per entitats comercials amb la intenció de fer-se publicitat o vendre productes. Aquestes pàgines acostumen a estar sota un domini .com (comercial).

c) Pàgines de notícies (*news Web pages*): tenen com a objectiu proporcionar informació molt actual. Sovint estan relacionades amb agències de notícies o altres entitats vinculades amb el món de la premsa. Aquestes pàgines acostumen a estar sota un domini .com (comercial).

d) Pàgines informatives (*informational Web pages*): tenen com a objectiu presentar informació objectiva. Aquestes pàgines acostumen a estar sota un domini .edu (institució educativa) o .org (entitat governamental).

e) Pàgines personals (*personal Web pages*): publicades per individus que poden estar vinculats a una organització o no. Aquestes pàgines es poden trobar sota qualsevol domini, i es freqüent que a la URL aparegui el signe ~, que indica que es tracta d'una pàgina personal.

Aquesta classificació és prou exhaustiva per incloure tots els recursos d'interès documental, tant per al món acadèmic com per als traductors[5]. Això no obstant, aquesta tipologia únicament inclou recursos textuals dissenyats per ser llegits, i no preveu recursos interactius que tenen com a objectiu un altre que no sigui l'obtenció d'un text (com podria ser un joc en línia).

Un cop definits els tipus de recursos que hom es pot trobar el *World Wide Web*, cal establir els criteris que permetran determinar si, dins de la seva categoria, un recurs és de qualitat o no. A aquesta tasca, la de dictar criteris sobre la qualitat dels recursos textuals del web, hi han dedicat sobretot entitats relacionades amb serveis d'informació, tant públics com privats, des de biblioteques fins a consultores. Aquestes entitats han posat en relleu el fet que, si bé és cert que hi ha molta trivialitat i molts recursos de baixa qualitat a la xarxa, no és menys cert que també hi ha (i hi haurà) molta informació d'alta qualitat que no es pot menysprear deixant-la fora dels índexs i bases de dades acadèmics o professionals, i que destriar els recursos de qualitat dels que no ho són facilita la tasca documental als usuaris del servei. De fet, molts d'aquests serveis ofereixen en línia el seu catàleg de recursos textuals, per la qual cosa acaben funcionant com un portal d'accés a recursos de qualitat, sovint a partir d'una taxonomia.

Aquest és el cas de la consultora The Argus Clearinghouse[6], la qual, a més de classificar els recursos en funció del tema o el format, també en determinava la qualitat en funció dels criteris següents:

- Tipus de recurs: es pressuposa una major qualitat a una pàgina informativa que a una de personal.

[5] La tipologia d'Alexandre i Tate s'adoptarà per classificar els corpus monolingües especialtizats objecte d'estudi d'aquesta recerca, tal com es veurà a l'apartat 2 "La classificació dels recursos textuals digitals especialitzats dedicats als Leònids" del capítol 4, pàgina 197.

[6] The Argus Clearinghouse: http://www.clearinghouse.net. Ha deixat d'actualitzar les seves dades des del gener de 2002.

- Dificultat en l'ús del recurs: el seu disseny gràfic, si resulta fàcil de llegir (des d'un punt de vista ergonòmic) o si utilitza els recursos gràfics de manera apropiada.
- Autoria i fiabilitat: si l'autor està capacitat per emetre informació de qualitat en aquell àmbit.

Però, sens dubte, són les biblioteques universitàries les entitats que més avenços han aportat en l'àmbit de l'establiment de criteris de qualitat aplicats als recursos textuals. La mateixa tipologia de recursos textuals proposada per Alexander i Tate (2001) va acompanyada de tota una sèrie de criteris que permeten establir la qualitat del recurs segons la categoria a què pertany. Els criteris són els següents:

1. **Autoria** (*authority*): es valora tant la presència de l'autor com d'un responsable secundari que legitimitzi la informació (si no és una plana personal). També es valora que faciliti la posada en contacte amb l'autor o responsable (sobretot si es una pàgina informativa o publicitària), o que tingui copyright (si és una pàgina de notícies).
2. **Precisió** (*accuracy*): si permet verificar les dades aportant una llista de recursos addicionals, si els continguts han estat revisats (procés editorial), o si hi ha faltes d'ortografia.
3. **Objectivitat** (*objectivity*): si hi consta cap tendència o inclinació de qualsevol mena del responsable (sobretot en els recursos de suport i defensa i en els personals), si inclou publicitat i es diferencia clarament del cos del recurs, o si se separa apropiadament la informació de l'opinió (sobretot en recursos de notícies).
4. **Actualitat** (*currency*): si hi consta la data de creació i/o revisió del recurs, la freqüència amb què s'actualitza o altres referències temporals.
5. **Abast** (*coverage*): si hi ha cap indicació sobre si la pàgina està acabada o no, si queda clar el tema que es tracta, si no es tracten aspectes importants del tema o si inclou enllaços cap a pàgines que permeten completar la informació.

Amb una intenció més exhaustiva que el conjunt de criteris que s'acaba de presentar, Codina recull tota una sèrie de paràmetres de qualitat englobats en quatre grans categories:

1. Contingut:
    a. Factors relatius a la qualitat: rigor en el tractament del tema, exhaustivitat, actualització de la informació, procés d'edició, sistematització (estructural i metodològica del contingut), interès intrínsec i originalitat de la informació.
    b. Factors relatius a la quantitat: superació del llindar de la trivialitat i aportació d'una cobertura relativament rellevant del tema.
2. Autoria:
    a. Solvència o adequació de l'autor o la institució responsables del recurs.
    b. Referències explícites a l'autoria.
    c. Esperança de validesa del recurs.
3. Ergonomia: disseny adequat al contingut que no dificulta la lectura o consulta del recurs.
4. Representació de la informació: presència d'eines que faciliten la navegació interna de manera estructural o conceptual.

En major o menor mesura, tant els criteris previstos per Alexandre i Tate com els que Codina proposa es repeteixen al llarg de tota la bibliografia dedicada a aquest tema, generada principalment, com ja s'ha assenyalat, en el marc de serveis d'informació públics. En conjunt, i a mode de sumari, els criteris que generalment es tenen en compte són els següents:

1. **Autoria**: tant si hi consten autor i responsable secundari com si estan autoritzats per publicar sobre el tema de què parlen. Aquest és un aspecte fonamental a l'hora d'establir la fiabilitat del recurs.

2. **Actualitat**: no només el fet que el recurs reculli informació tan actual com sigui possible, sinó també que hi consti la data de creació o actualització, de manera que es pugui comprovar que la informació era completa i vàlida en el moment de ser publicada.
3. **Precisió**: en el contingut i també en la forma, atès que un text sense errades ortogràfiques o d'altra mena denota un contingut acurat.
4. **Tractament del contingut**: en quina mesura es tracta un tema determinat (el seu abast), de quina manera (objectivament o subjectivament, és a dir, com a informació o com a opinió) i de quina forma (si inclou altres morfologies de la informació a més de la textual i ho fa perquè la informació que s'aporta ho requereix).
5. **Originalitat**: demostrar que la informació tractada és original o, si no ho és, indicar les fonts de referència.
6. **Propòsit**: deixar palesa la inclinació o tendència de qualsevol mena de l'autor o responsable per tal de facilitar la interpretació del contingut del recurs.
7. **Enllaços amb altres recursos**: permetre la validació del contingut del recurs proporcionant enllaços a altres recursos, i que aquests siguin de qualitat. També es valora el fet que no siguin una mera llista d'accessos a altres pàgines, sinó que estiguin comentats de manera que el lector conegui breument el contingut de la pàgina destí de cada enllaç abans d'accedir-hi.
8. **Ergonomia**: el disseny no ha de dificultar la consulta del recurs, i ha de facilitar la navegació interna (mitjançant un menú i/o un mapa del recurs).
9. **Entorn informàtic**: si requereix components (de *software* o *hardware*) addicionals per consultar el recurs, si la seva presència aporta un valor afegit al recurs i si aquests components són habituals i fàcils d'utilitzar o no.
10. **Citació**: si el recurs és citat a altres recursos, cosa que denota qualitat.
11. **Receptor**: quin és el receptor que l'emissor tenia en ment en elaborar el recurs. (Aquest és un criteri difícil d'establir, ja que en la pràctica qualsevol

internauta pot ser el receptor de qualsevol recurs textual, encara que aquest no hagués estat dissenyat pensant en el seu perfil.)[7]

Generalment els criteris, presentats com a llista d'aspectes que cal tenir en compte en avaluar un recurs textual digital, van acompanyats per una bateria de preguntes que permeten establir el nivell de qualitat del recurs, de manera que s'acaben convertint en tests que els recursos han de superar per tal de ser inclosos a les bases de dades dels serveis d'informació. Fonamentalment, les preguntes que es proposen als diferents tests es responen analitzant el component lingüístic del recurs, el text, com també altres components extralingüístics, com poden ser els formats utilitzats per a cada element present al recurs, la seva adreça URL o l'origen de l'autor.

Amb tot aquest conjunt de criteris s'intenta compensar l'absència dels condicionants del procés editorial durant l'elaboració dels recursos, i en la pràctica permet destriar els que haurien pogut passar pel procés editorial i ser publicats, els de qualitat, dels que mai no ho haurien estat.

En el cas del traductor, usuari potencial dels recursos textuals digitals com a font de documentació per a la presa de decisions emmarcades en una traducció determinada, el recurs de qualitat en cada moment és aquell que, a més de complir tots els requisits indicats anteriorment, s'adequa a la situació comunicativa del seu encàrrec de traducció; dit amb altres paraules, el recurs de qualitat ha de ser sobretot pertinent ateses les necessitats informatives del traductor. Per aquest motiu, recursos que un servei d'informació podria considerar de qualitat insuficient per ser inclosos en la seva base de dades, poden ser els més adequats a les necessitats informatives del traductor i, per tant, de suficient qualitat per recolzar una decisió de traducció en el seu contingut. Això no obstant, conèixer aquests criteris de qualitat li poden permetre seguir unes estratègies de documentació concretes en funció de la seva necessitat informativa, amb l'objectiu d'arribar al recurs o recursos que més s'hi adeqüin.

[7] La compilació dels criteris més rellevants i la seva ordenació per importància ha estat realitzada per l'autora.

Un cop revisades les característiques tant formals i estructurals com les relatives al contingut dels recursos web, i atès que l'entorn web és tan immens com caòtic, a continuació es descriuran els diferents mecanismes de recuperació d'informació en aquest entorn, començant pels seus fonaments, les tècniques i les metodologies.

1.2. La recuperació d'informació

El concepte de **recuperació d'informació** (RI) és relativament nou. Tal com la seva denominació deixa albirar, consisteix a recuperar documents d'un fons documental en funció d'una necessitat informativa determinada. Aquest procés implica l'ús d'ordinadors, ja que s'entén que tant el fons documental, o una representació (en el cas de les bases de dades referencials), com el procés de cerca es fa de manera informatitzada.

L'objectiu de tot sistema d'RI és proporcionar els documents adequats a la consulta d'un usuari entre els documents del seu fons. Per aquest motiu, tant la manera d'organitzar el fons que el nodreix com d'arribar-hi esdevenen fonamentals per al seu bon funcionament.

Per tal de poder recuperar els textos que responen a una necessitat informativa determinada, un sistema d'RI ha de comptar amb: (1) un subsistema de representació de la informació continguda als textos del seu fons (l'equivalent a les fitxes que identifiquen els llibres d'una biblioteca), (2) un subsistema que permeti realitzar consultes als usuaris i (3) un subsistema que creui les consultes amb la representació dels documents per tal de trobar aquells que l'usuari necessita (Codina, 2002: 5; Chowdhuri, 1999: 5). Es tracta, per tant, d'un sistema complex que emmagatzema i recupera textos i que no informa l'usuari sobre una matèria concreta, sinó que es limita a indicar-li si existeixen textos relacionats amb la seva consulta o no (Chowdhuri, 1999: 1).

Generalment, un cop localitzat el fons documental que formarà part d'un sistema d'RI, s'observa els documents que en formen part i se'n representa el contingut, de manera

que en fer-hi una consulta el sistema cerca a la representació dels documents aquells que més s'hi escauen.

El més habitual és representar el contingut dels documents amb paraules clau que el resumeixen, com es veurà en el proper apartat. En formular una consulta, el sistema d'RI creua les paraules utilitzades per l'usuari amb els descriptors del contingut dels textos i en selecciona els més pertinents.

Sovint, la tasca de representació del contingut dels documents es fa de manera semiautomatitzada tot reduint el contingut a les paraules o expressions més significatives del text. Per tal d'incrementar l'efectivitat dels sistemes d'RI, tant si compten amb un sistema d'anàlisi i representació automatitzat com si no, s'acostuma a limitar els possibles descriptors que poden identificar cadascun dels documents, generalment establint una llista tancada prèviament. Es tracta, per tant, d'exercir un cert control sobre el vocabulari utilitzat en la fase d'anàlisi i representació del contingut del fons documental, de manera que, si les consultes s'expressen utilitzant el mateix vocabulari controlat amb què el sistema identifica el contingut dels textos, l'usuari obtindrà el màxim rendiment del sistema.

A continuació es descriuen els diferents sistemes d'anàlisi i representació del contingut d'un fons documental i les seves implicacions per tal d'analitzar més endavant els diferents sistemes d'RI que s'han implementat en l'entorn web.

1.2.1. La representació de la informació

El gran potencial dels sistemes d'RI rau en la capacitat de creuar les necessitats d'informació dels usuaris expressades de forma lingüística amb els descriptors del contingut del fons documental. El contingut de cada document es representa de manera que faciliti aquest encreuament amb les consultes dels usuaris. Aquesta representació es pot realitzar a partir d'una indexació o descripció característica, mitjançant la qual s'identifica la matèria dels documents a través d'uns termes característics que permeten la seva posterior recuperació; d'altra banda, també es pot

realitzar un resum o descripció substancial, que permet condensar i conèixer amb profunditat el contingut dels documents.

El procés de representació del contingut dels documents en un sistema d'RI es basa generalment en la **indexació**, procés que consisteix en la identificació i descripció dels conceptes informatius expressats de manera explícita o implícita al contingut dels documents, i a representar-los mitjançant els denominats termes d'indexació. A l'hora de realitzar aquesta operació, els documents es poden indexar per matèries, per conceptes o per paraules clau. La indexació per matèries és una operació sintètica, que consisteix a identificar el tema principal del document i representar-lo, generalment a través d'un únic terme d'indexació. La indexació per conceptes és una operació analítica, mitjançant la qual s'identifiquen els conceptes que formen el contingut del document i es representa a través d'una sèrie de descriptors que, en general, s'estableixen prèviament. La indexació per paraules clau és una operació que analitza el contingut dels documents a través de les paraules significatives que contenen el títol, el resum o el text complet.

Generalment, a Internet, els termes d'indexació utilitzats són paraules o descriptors. Les paraules clau són aquelles paraules o grups de paraules seleccionades de forma automàtica del títol, el resum o el text complet d'un document, i representen el seu contingut alhora que permeten la seva recuperació. D'altra banda, els descriptors són termes d'indexació assignats pels analistes, fruit d'alguna de les operacions intel·lectuals que implica el procés d'indexació: examen del document, identificació dels conceptes i traducció a un llenguatge documental.

La indexació de documents utilitzant llistats de descriptors requereix un esforç intel·lectual de comprensió del contingut dels documents i la seva síntesi en descriptors, per la qual cosa no s'acostuma a fer de manera automatitzada (Codina, 2002). En la **indexació automàtica** de documents, la intervenció de l'ordinador no consisteix a analitzar el contingut dels documents, sinó a resumir-lo gràcies a l'anàlisi del lèxic de cadascun d'aquests, puix que avui dia els ordinadors no podrien dur a terme aquesta tasca, almenys sobre text pur sense cap mena de marques o etiquetes.

Habitualment, un sistema d'indexació automàtica duu a terme les operacions següents:

1. Crea una llista amb les paraules del document.
2. Elimina les paraules més comunes (paraules funcionals o buides).
3. Ordena la llista resultant en funció de la freqüència d'aparició de les paraules al document.
4. Escull com a descriptors totes aquelles paraules que superen una freqüència mínima. (Chowdhury, 1999: 88).

La llista de descriptors resultant s'inclou en un **arxiu invers** que conté tots els descriptors utilitzats ordenats alfabèticament. La base de dades d'un sistema d'RI basat en un arxiu invertit està dividida en dos arxius: el primer conté els registres dels textos que componen el fons documental (l'equivalent a un catàleg de biblioteca) i el segon, l'arxiu invertit, conté tots els descriptors indexats, i cadascun d'ells hi té associat un identificador del text o textos en els quals hi apareix, i fins i tot la posició que hi ocupa. D'aquesta manera, cada cop que es formula una consulta amb una sèrie de descriptors, el sistema recupera tots aquells documents que els contenen.

De tot el que hem dit fins ara es desprèn que la indexació automàtica de documents només es pot dur a terme sobre un fons documental en format digital, com ara els documents que conformen el *World Wide Web*. Els sistemes d'RI sobre fons documentals digitalitzats permeten la ràpida localització de textos, que sovint també proporcionen, com es veurà a l'apartat següent.

A més dels sistemes d'indexació manual (la metodologia clàssica d'anàlisi i representació del contingut de documents) i dels d'indexació automàtica (la implementació informàtica dels primers), s'han desenvolupat altres metodologies amb els mateixos objectius que han tingut una gran repercussió en l'entorn WWW, com són ara la indexació amb robots o agents i la indexació a partir de metainformació (vegeu l'apartat 1.3 "Els sistemes de recuperació d'informació al *World Wide Web*", pàgina 144).

1.2.2. La selecció i obtenció d'informació

Tal com s'indicava a l'apartat anterior, els sistemes d'RI recuperen documents tot creuant l'expressió d'una necessitat informativa realitzada per l'usuari amb els descriptors que representen el contingut del fons documental. Aquest apartat està dedicat a les diferents metodologies d'interrogació informatitzada d'aquests sistemes, que seran les més utilitzades pels sistemes d'RI del WWW. En concret s'analitzaran els quatre models següents: el model Booleà, amb més detall, i més sumàriament el model vectorial, el probabilístic i la cerca per patrons (*pattern matching*).

El **model Booleà** compara els descriptors utilitzats en la representació del contingut del fons documental amb els de la consulta, tot establint quins documents són rellevants i quins no ho són. Les consultes s'expressen utilitzant els operadors de la lògica Booleana: AND (que recupera documents que continguin un i altre terme indexat), OR (que recupera documents que continguin un i/o l'altre terme indexat) i NOT (que recupera documents que no continguin un terme indexat). Amb aquests tres operadors es poden construir consultes amb un grau de complexitat més o menys alt en funció dels coneixements de la lògica de l'usuari. Utilitzar únicament la lògica AND, o coincidència exacta, acostuma a restringir molt el volum de la resposta facilitada pel sistema, mentre que la lògica OR, o coincidència parcial, proporciona respostes més àmplies.

En qualsevol cas, en els sistemes Booleans clàssics la resposta està formada per documents que són rellevants, en funció d'una lògica o l'altra, sense cap mena de gradació: el sistema només discrimina entre documents rellevants i els que no ho són (que no compleixen els condicionants expressats a la consulta), però no és capaç de determinar entre els rellevants quin document ho és en un major grau i quin d'ells ho és en menor grau.

El **model vectorial** es basa en la comparació de la representació tant dels documents com de la consulta en conjunts de termes per calcular-hi el grau de similitud. Al contrari que amb el model Booleà, amb el vectorial no s'obtenen documents rellevants o no, sinó que s'obté una llista de documents ordenats en funció de la similitud entre la seva representació i la consulta realitzada. Per tal de sospesar les dades de la manera més

pertinent possible, a l'hora de realitzar el càlcul de la similitud, basat en un algorisme, se li pot donar un major pes a uns descriptors determinats en funció, per exemple, del lloc que ocupen al document (formen part del títol o apareixen en negreta), o de les necessitats d'informació de l'usuari; així doncs, la rellevància d'un document rau en si conté una sèrie de descriptors i en quina posició hi apareixen. Generalment, però, aquest sistema de cerca pren els descriptors com a unitats independents, aspecte que sovint es considera un dels seus inconvenients més importants.

Actualment, gairebé tots els sistemes d'RI que apliquen el model Booleà també permeten ordenar els resultats d'una cerca en funció de la seva rellevància, que es calcula valorant el pes de cada terme cercat dins dels documents recuperats. Així doncs, un cop el sistema separa els documents que són pertinents donada una cerca determinada dels que no ho són, ordena els primers atorgant valors a cadascuna de les ocurrències dels termes cercats per tal de calcular el grau de rellevància i mostrar els resultats de la cerca en forma de llista descendent en funció de la rellevància de cadascun dels documents.

El **model probabilístic** es basa en la interacció entre el sistema i l'usuari. Aquest últim descriu aproximadament les característiques temàtiques i/o formals dels documents que vol recuperar, el sistema li proposa una resposta, i en funció de la resposta l'usuari va refinant la cerca fins que arriba a la resposta ideal. Un cop s'arriba a aquest document ideal, el sistema ordena el fons documental en funció del grau de similitud amb el document ideal. Aquest model assumeix que la probabilitat de rellevància únicament depèn de la representació de la consulta i del document, sense tenir en compte altres factors i pressuposant que sempre hi haurà un document ideal per a cada consulta (Chowdhury, 1999: 31). El gran inconvenient de l'aplicació d'aquest model és el procés d'identificació del document ideal que, a més de llarg, pot resultar difícil d'implementar.

El **model de cerca basat en patrons** (*pattern matching*) no pren com a unitat de cerca el descriptor, sinó els elements morfosintàctics que el configuren. D'aquesta manera, la cerca es pot realitzar a partir de paraules o de parts de paraules, com ara afixos, o bé cadenes de caràcters. El sistema cerca documents amb cadenes de

caràcters idèntiques o similars a la de la consulta, és a dir, que teòricament permet recuperar documents a partir de paraules amb errors tipogràfics (Chowdhury, 1999: 105). Amb aquests models, la representació del contingut mitjançant arxius invertits resulta molt més complexa, ja que la unitat a tenir en compte és el caràcter i el seu context, és a dir, els caràcters que l'envolten, a més d'unitats com afixos o paraules.

A l'hora de valorar el funcionament i l'efectivitat d'un sistema d'RI es tenen en compte principalment dos aspectes: la taxa de **recordació**[8] (*recall*) i la de **precisió** (*precision*) (Baeza-Yates i Riveiro-Net, 1999; Chowdhury, 1999; Codina, 2002; Lancaster, 1968). La taxa de recordació d'un sistema d'RI és la proporció de tot el material rellevant del fons documental que el sistema és capaç de recuperar; permet esbrinar quin percentatge de documents (temàticament) rellevants no ha estat tingut en compte en la resposta que el sistema ofereix a una consulta determinada. La taxa de precisió, d'altra banda, indica la proporció de documents rellevants amb relació a tots els documents recuperats; per tant indica quin percentatge dels documents recollits a la resposta no són (temàticament) rellevants tenint en compte la consulta per a la qual han estat recuperats.

Aquests dos paràmetres fonamentalment permeten analitzar l'efectivitat d'un sistema d'RI i es veuen afavorits com més exhaustiu és el procés d'indexació de la informació, tot i que també estan condicionats per la manera d'expressar la consulta. Generalment, en fer una consulta molt àmplia, la taxa de recuperació acostuma a ser alta, és a dir, el sistema recupera un percentatge molt alt dels documents rellevants. La resposta, però, també acostuma a incloure molts documents que no són pertinents, per la qual cosa la taxa de precisió se situa a nivells molt baixos. Per exemple, si se cerquen documents que continguin les paraules *pluja* i *Leònids*, sense cap altra restricció, es recuperarà pràcticament tots els documents que parlin de *pluja de Leònids*, però també els que parlin únicament de *pluja* (com ara informes meteorològics), que no serien pertinents.

Per tal de millorar la taxa de precisió cal realitzar cerques molt més restrictives. Així

[8] Recordació és la traducció del terme *recall* que proposa Codina en el seu article "Fonaments de teoria de recuperació d'informació" (Codina, 2002).

doncs, si en lloc de cercar paraules individuals se cerca l'expressió *pluja de Leònids*, la taxa de precisió augmentarà, puix que el resultat no inclourà soroll com ara parts meteorològics. Tanmateix, el sistema no podria recuperar els documents que continguessin la cadena *pluja denominada dels Leònids*, cosa que sí que hauria pogut fer en el primer exemple, per la qual cosa la taxa de recuperació disminueix. Com es pot desprendre, com més alta és la taxa de recuperació, més baixa resulta la de precisió, i viceversa.

La taxa de recuperació i la de precisió són, doncs, valors pràcticament inversament proporcionals. Tot i que en dissenyar els diferents sistemes d'RI els experts recomanen mantenir un equilibri entre tots dos valors, i que tant la taxa de recuperació com la de precisió es trobin entre el 50% i el 60% (Chowdhury, 1999: 71), en formular la seva consulta l'usuari també pot condicionar aquests valors. En alguns casos, l'usuari pot estar més interessat a aconseguir tants documents com sigui possible, primant la quantitat per sobre de la qualitat, per la qual cosa realitzarà una **cerca amb una taxa de recordació alta** (*high recall search*); en altres casos, l'usuari pot preferir recuperar menys documents, però tots ells rellevants, primant per tant la qualitat per sobre de la quantitat, per la qual cosa realitzarà una **cerca amb una taxa de precisió alta** (*high precision search*). Sovint, sobretot en l'entorn web, l'usuari d'un sistema d'RI en té prou amb recuperar fins i tot un sol document, sempre que aquest cobreixi les seves necessitats informatives, encara que el fons documental en contingui més, de documents rellevants; en aquest cas l'usuari realitza una **cerca breu** o *brief search* (Chowdhury, 1999: 159).

En qualsevol cas, l'usuari ha de conèixer certs aspectes del sistema d'RI que fa servir, com ara el model teòric en què basa el sistema de recuperació de documents o les característiques fonamentals del seu sistema d'indexació, per tal de treure'n tot el profit i poder formular cerques que obeeixen al seu objectiu. En paraules de Chowdhury, "developing a good search strategy requires knowledge about the nature and organization of target database(s) and also the exact needs of the user" (1999: 158).

1.3. Els sistemes de recuperació d'informació al *World Wide Web*

Un cop revisades les característiques essencials de tot sistema d'RI, en aquest apartat ens centrarem en els sistemes d'RI del *World Wide Web*: els coneguts de manera genèrica com a **cercadors**. En un primer moment analitzarem el funcionament d'aquest sistemes, i a continuació, per tal d'establir la cobertura dels seu fons respecte del total del *World Wide Web*, descriurem les característiques formals del document típic que pot formar part del seu fons documental i del que mai no hi podrà formar part segons les característiques actuals dels cercadors. Finalment, revisarem nous possibles sistemes d'RI que s'estan posant en pràctica.

En un marc hipertextual com l'entorn web, la manera paradigmàtica d'accedir a la informació és navegar, seguint els vincles que permeten aprofundir en els nodes d'un document hipertextual o que els enllacen amb nodes d'un altre document. Enfront d'aquesta metodologia d'obtenció d'informació, la navegació, es troba la que Rovira denomina la recuperació per interrogació: la consulta de sistemes d'RI que faciliten l'accés als llocs web (2002: 114). Per tots és sabut que no hi ha cap catàleg de llocs web, cap llistat exhaustiu dels documents hipertextuals que en formen part. És per això que sovint s'afirma que Internet es troba en un estat proper al caos pel que fa a la manera d'accedir-hi i per com està organitzada (Chowdhury, 1999: 215; Woodward, 1995), ja que per trobar un document del qual no en coneixem l'adreça només podem recórrer a sistemes d'RI la base de dades dels quals és limitada, i que sovint només es pot interrogar a partir de paraules.

1.3.1. Els cercadors segons el seu funcionament

De cercadors, n'hi ha de diferents tipus i s'acostumen a classificar en funció de criteris diversos, com ara si són especialitzats o genèrics, o si són regionals o globals. De tota manera, però, el criteri diferenciador més habitual acostuma a ser el següent: si permeten un accés a la informació tot navegant per categories o bé realitzant una consulta amb paraules clau. En el cas dels primers, els denominats **directoris** o

índexs temàtics, la informació acostuma a estar recollida per especialistes que únicament inclouen en cadascuna de les categories temàtiques documents amb informació rellevant, és a dir, documents de qualitat. Els segons, els denominats **motors de cerca**, permeten interrogar bases de dades recollides automàticament, sense intervenció humana o amb una intervenció mínima.

Avui dia, la majoria de directoris també permeten interrogar la seva base de dades amb consultes directes, i molts motors de cerca alhora també proporcionen un accés temàtic pel qual l'usuari pot navegar. La gran diferència entre tots dos sistemes d'RI és com recullen les seves dades. A més, existeixen altres sistemes d'RI, com ara els metacercadors, que es descriuran en profunditat al final d'aquest apartat, ja que es consideren evolucions dels dos primers sense que hi aportin cap innovació significativa pel que fa a la recollida de la informació o al seu accés.

a) Els directoris o índexs temàtics

Els directoris són compilacions de recursos en línia realitzades per persones, experts en documentació i biblioteconomia i/o en l'àmbit temàtic de què tracta el recurs o que cobreix el directori. Aquest procés, que pot resultar lent o poc productiu en comparació a la indexació automàtica dels motors de cerca, permet construir bases de dades únicament amb recursos de qualitat, ja que han estat avaluats i classificats per experts, i indexats en la categoria pertinent (Mas i Sallas, 1999: 28).

La tasca de l'indexador es basa en l'anàlisi de cada recurs textual. Generalment, l'indexador arriba a conèixer l'existència del recurs ja sigui perquè el seu autor o webmaster el dóna a conèixer al directori o perquè a aquest nou recurs hi van a parar enllaços des de pàgines que ja han estat incloses a la base de dades del directori. En primer lloc, detecta els termes que descriuen el contingut del text, que poden ser unitats d'una sola paraula o de més d'una. A continuació, i aplicant la metodologia d'indexació amb un llenguatge controlat, hi assigna els descriptors pertinents que estan relacionats amb cadascuna de les categories del directori. D'aquesta manera, el document pot quedar indexat per descriptors que no conté explícitament i viceversa,

alhora que la llista de descriptors o tesaurus que faci servir el directori es pot veure ampliada o completada en funció de les opinions dels experts i de la informació que analitzen, puix no es tracta d'una llista definitivament tancada.

Una altra de les característiques dels directoris es que s'organitzen en forma de classificacions temàtiques jeràrquiques, de manera que cada cop que s'accedeix a un apartat se'n mostren els subapartats, i en cadascun d'aquests segons els documents que hi estan classificats, seguint un esquema de classes i subclasses. Cada directori acostuma a ordenar els documents en una sèrie de categories pròpies, que comercialment els dóna resultats, però que no responen a cap taxonomia jeràrquica representativa del coneixement humà en tota la seva complexitat, com ara la classificació decimal universal. Aquest és el cas del directori més conegut i visitat, Yahoo![9] o de la plana inicial de qualsevol portal genèric (que també funcionen a mode de directori, tot i que sovint inclouen recursos que paguen per constar-hi i que, per tant, s'hi anuncien). En partir de categories temàtiques creades arbitràriament, algunes es poden arribar a encavalcar, o el que es pitjor, poden quedar llacunes de coneixement per cobrir. Aquest, i el fet que cada directori té una base de dades pròpia, diferent de la dels altres directoris, fa que en consultar-los s'obtinguin resultats diferents. Tot i així, també existeixen lloables excepcions que recorren a classificacions temàtiques àmpliament contrastades i que fins i tot s'utilitzen en serveis d'informació convencionals (biblioteques o arxius), com ara BUBL[10] o SOSIG[11].

Es calcula que les bases de dades de tots els directoris juntes cobreixen aproximadament un 1% del total del *World Wide Web* (Baeza-Yates i Ribeiro-Net, 1999: 384). Tot i així, els directoris són considerats una molt bona eina per iniciar una

[9] Yahoo!: http://www.yahoo.com. Ofereix interfícies en diferents idiomes, també en castellà i català.

[10] BUBL: http://www.bubl.ac.uk. Servei d'informació del Centre for digital Library Research de la Universitat d'Strathclyde, Glasgow, anomenat d'aquesta manera pel seu **BU**lletin **B**oard for **L**ibraries, i que principalment recull informació relacionada amb l'àmbit de la documentació i la biblioteconomia.

[11] SOSIG: http://www.sosig.ac.uk. **SO**cial **S**cience **I**nformation **G**ateway. Forma part de la UK Resource Discovery Network, i únicament recull recursos d'alta qualitat, vàlids en el món acadèmic i, principalment, en llengua anglesa.

cerca en un àmbit temàtic que no es coneix, ja que proporcionen els recursos més importats de cada tema (Codina, 2000b: 167) i les seves respostes sempre són pertinents, sempre que la seva base de dades s'actualitza periòdicament per tal de tenir els seus registres al dia.

L'exemple paradigmàtic de directori és Yahoo!, que és el cercador més utilitzat pels internautes (Sullivan 2001b). Tot i que també permet la recuperació d'informació a partir de la consulta directa amb paraules clau, el seu punt fort és l'accés a la informació per navegació a través de les seves categories, que són:

- Art i cultura: literatura, museus, teatre, etc.
- Internet i ordinadors: WWW, programes, xat, etc.
- Ciència i tecnologia: animals, ecologia, enginyeria, etc.
- Materials de consulta: biblioteques, diccionaris, traductors,etc.
- Ciències socials: història, lingüística, psicologia, etc.
- Mitjans de comunicació: notícies, diaris, ràdio, TV, etc.
- Economia i negocis: empreses, immobiliàries, ocupació, etc.
- Política i govern: ajuntaments, dret, política, etc.
- Ensenyament i formació: escoles, selectivitat, universitats, etc.
- Salut: hospitals, malalties, medicina, etc.
- Espectacles i diversió: postals, cinema, música, genial!, etc.
- Societat: festes, gastronomia, per a nens, etc.
- Esports i lleure: futbol, jocs, turisme, etc..
- Zones geogràfiques: províncies, poblacions.[12]

En analitzar aquesta classificació temàtica es veu clarament com alguns àmbits de la vida quotidiana es tracten amb molt de detall, mentre que altres, importants en la història del coneixement humà però potser no tan rellevants en el dia a dia de qualsevol ciutadà del món occidental, com ara la filosofia o la teologia, ni tan sols hi apareixen en les primeres subcategories. En aquest cas, a més, també s'observa com la

[12] Categories extretes de la interfície de Yahoo! en català, el 7 d'abril de 2003.

classificació ha estat adaptada a les característiques culturals i socials dels potencials usuaris, en aquest cas un internauta de parla catalana; per aquest motiu s'inclouen subcategories com per exemple "Selectivitat" en la categoria d'"Ensenyament i formació", o la categoria denominada "Zones geogràfiques" que està subdividida en "Províncies" i "Poblacions", apartats tots que clarament no són ni universals ni exportables tal com estan expressats, sinó que responen a la realitat social i cultural del nostre país.

En els directoris, quan també permeten realitzar cerques directes, es pot obtenir informació per **navegació i després cerca** (*browse and then search*) o bé per **cerca i després navegació** (*search and then browse*). En el primer cas, un cop s'ha començat a navegar tot aprofundint en una categoria determinada, es pot realitzar una cerca amb paraules clau dins de la categoria en què l'usuari es troba. En el segon cas, al contrari, l'usuari pot realitzar una consulta amb paraules clau que li portin a una categoria o subcategoria de la classificació temàtica del directori i seguir navegant a partir d'aquell punt.

A més dels directoris d'àmbit genèric, també n'hi ha d'especialitzats: dedicats únicament a un àmbit determinat del coneixement humà i/o a una zona geogràfica específica. Es tracta de guies recollides per especialistes i ordenades en categories (Tyner, 2002), que s'actualitzen i permeten la navegació com si d'un directori genèric és tractés, i no s'han de confondre amb els índexs que sovint formen part dels llocs web, recollits en la majoria dels casos sense ànim de ser exhaustius ni precisos i que no acostumen a permetre la navegació.

Els registres obtinguts amb una cerca en un directori, a més de l'adreça d'accés a cadascun dels llocs web proposats, també hi acostuma a constar un petit resum o un extracte del lloc, així com el darrer cop que ha estat revisat pels responsables de la base de dades del directori. A vegades també hi consta un marca de ponderació, sobretot si es tracta d'un directori especialitzat, o d'un directori que permet als usuaris votar els llocs que consideren de millor qualitat. Generalment es recullen els documents hipertextuals com a unitat, no s'indexa cadascun dels nodes o pàgines per separat, de

manera que els resultats acostumen a apuntar cap a la plana inicial dels documents proposats.

b) Els motors de cerca

El funcionament dels motors de cerca es basa en l'ús d'índexs invertits creats automàticament com a base de dades. Al contrari que amb els directoris, els motors de cerca, tot i que en alguns casos ofereixen un accés temàtic a la informació, faciliten l'accés als documents recollits a la seva base de dades mitjançant un sistema de consultes directes.

Fonamentalment, el motor de cerca està format per tres elements:

1. Un programa o robot que cerca informació al web.
2. Un índex o base de dades on es recull la informació.
3. Un sistema que cerca a la base de dades els resultats més adequats a cada consulta (Sullivan, 2001a).

El motor de cerca és, per tant, un sistema d'RI pròpiament dit, tot i que alguns presenten variacions d'aquest esquema per tal de millorar la seva efectivitat en cada circumstància.

El programa que localitza i permet la indexació dels recursos que troba es denomina genèricament **robot** o **agent**, tot i que també rep altres denominacions d'acord amb el seu funcionament, com ara *crawler*, *spider*, *walker* o *wanderer*. Aquests programes surten de llocs web que el sistema ja coneix, i segueixen els enllaços hipertextuals externs que contenen cercant documents nous que no formin part de la base de dades del motor. Quan en troba un, n'envia una còpia al motor per tal que l'indexi a la base de dades a partir de les seves paraules més freqüents o més representatives. Com que en la fase d'anàlisi i indexació de la informació no es fa cap activitat intel·lectual, en paraules de Codina (2000b), sinó que únicament es realitza en funció del l'anàlisi de freqüències d'aparició del lèxic que forma part de cada document, obtenir una classificació temàtica sense cap més tret o recurs addicional pot resultar utòpic.

Els robots només poden processar informació en format text, puix que a l'índex només s'hi troben paraules, i no poden identificar característiques d'un document com ara el seu context o la seva temàtica. L'heterogeneïtat formal regnant entre els documents hipertextuals que configuren el web tampoc no ajuda a interpretar-los o analitzar-los automàticament: si tots els recursos textuals d'unes característiques determinades continguessin uns trets o unes marques representatives, els robots les podrien interpretar i utilitzar-les per catalogar-los correctament (vegeu el subapartat d) "Altres sitemas de cerca" d'aquest apartat, pàgina 162). Una de les crítiques més habituals que s'acostuma a fer a aquesta manera d'aconseguir dades, a més de la baixa qualitat de les bases de dades que permeten crear en comparació a les dels directoris, rau en el fet que els robots són els responsables d'una gran part del trànsit d'Internet, i en ocasions poden arribar a saturar un directori amb les seves peticions de còpies de documents (Baeza-Yates i Ribeiro-Net, 1999: 374).

Durant el procés d'indexació automàtica dels documents web s'identifiquen les cadenes de caràcters discretes que formen part del text, les paraules, que es convertiran en termes d'indexació, i cadascun d'aquests termes s'assigna com a descriptor del document. Es tracta, en la majoria dels casos, d'una indexació en l'àmbit submorfològic, ja que no es realitza cap mena d'anàlisi morfològica, sintàctica o semàntica (Olvera, 1999); d'aquesta manera s'aconsegueix un sistema d'RI força flexible. En expressar la consulta, gairebé tots els motors de cerca permeten fer-ho amb operadors Booleans o sistemes similars que permeten formular consultes amb un cert grau de complexitat.

D'altra banda, en la fase de recuperació d'informació, el sistema, després de localitzar al seu índex els documents que compleixen els requisits expressats a la consulta, duu a terme una ponderació dels llocs trobats a partir d'un algorisme basat essencialment en el lloc que ocupi la paraula cercada a cada document i la seva freqüència d'aparició; és el que es coneix com a **mètode d'ubicació i freqüència** (*location/frequency method*) (Sullivan, 2001b): el sistema dóna prioritat a un document o un altre en la llista de resultats segons en quin lloc del document es troba la paraula cercada i el nombre de cops que s'hi repeteix. Tot i que cada motor de cerca fa servir un algorisme propi que

normalment no es dóna a conèixer amb detall, habitualment els criteris de ponderació ordenen els documents de la manera següent:

1. Aquells documents en què la paraula cercada figura al contingut del camp "Title" de l'encapçalament (*header*)[13].
2. Aquells documents en què la paraula cercada figura a l'inici del seu text: al títol o als primers paràgrafs.
3. Aquells documents en què la paraula apareix en més ocasions (freqüència d'aparició).

Es tracta, per tant, de criteris que avaluen únicament el contingut dels documents, però no de manera conceptual o morfosintàctica, sinó únicament com una cadena de caràcters interrompuda per espais en blanc o signes de puntuació.

A mesura que augmentava el volum del web, els motors van incorporar més i més informació al seu índex, sovint sense millorar el seu sistema de recuperació d'informació i actualitzant amb poca freqüència les seves dades. Per aquest motiu, els resultats acostumaven a ser molt extensos, tot i que de poca qualitat. Molts experts, referint-se als cercadors en general, els descrivien com a eines amb un baix nivell d'eficàcia:

> "Either they treated the Web pages into giant directories, or they played a numbers game, using computers to count matching words and phrases on millions of Web pages. The first approach was time-consuming and expensive; the second is faster but often imprecise." (Thottam, 2001)

Thottam descriu en primer lloc els directoris, les bases de dades dels quals eren i encara acostumen a ser compilades per experts, i en segon lloc els motors de cerca, que classifiquen i recuperen documents a partir de la presència o absència de les paraules que l'usuari cerca o descarta. Aquests primers motors de cerca realitzaven les

[13] El contingut dels camps que es troben a la part denominada encapçalament del document en format HTML es coneix com a *metadades* (*metadata*)

seves cerques sobre bases de dades relativament petites: AltaVista[14], per exemple, no superà el límit dels 250 milions de documents indexats fins a l'any 2000, tot i ser un dels primers motors de cerca públics del *World Wide Web*.

A partir de la segona meitat de la dècada dels noranta apareixen nous motors de cerca, com ara Google[15], que proporcionen uns resultats la precisió dels quals és sensiblement més alta que la dels seus predecessors. Aquesta millora rau en la implementació d'un nou criteri per establir la ponderació per rellevància dels documents: l'**anàlisi de citacions** (*citation analysis*). Segons aquest nou criteri, un cop localitzats els documents que responen a una consulta, el sistema calcula quants enllaços hipertextuals entre el conjunt de pàgines localitzades van a parar a cadascuna d'elles, com si de la votació del més popular es tractés (un enllaç amb el document X com a destí = un vot per a aquest document). D'aquesta manera s'aconsegueix basar el càlcul de rellevància amb què es mostraran els resultats en opinions fonamentades, és a dir, opinions expertes, ja que en publicar una pàgina sobre un tema determinat i proposar una sèrie d'enllaços com a pàgines addicionals, implícitament s'està donant validesa als recursos proposats. El fet és que, gràcies a aquest sistema, pràcticament s'ha aconseguit acabar amb el soroll en els primers 10 resultats d'una cerca extensa, sempre que la consulta hagi estat expressada adequadament.

Per tal de millorar la seva efectivitat, els motors de cerca també han perfeccionat el procés d'indexació de la informació: no indexen els documents que localitzen a partir de totes les paraules que contenen, sinó únicament d'aquelles que faciliten la discriminació entre documents rellevants i no rellevants en formular una consulta, per la qual cosa s'eliminen les paraules buides o funcionals i les que són tan habituals que pràcticament es troben a tots els textos (p. ex., les formes del present del verb ser) mitjançant una

[14] Altavista: http://www.altavista.com. Avui dia també ofereix un accés a la informació per categories, fonamentalment de tipus comercial, i ofereix interfícies en més de 15 idiomes adaptades a les característiques culturals de més de 20 països.

[15] Google: http://www.google.com. S'hi pot accedir mitjançant interfícies en més de 50 llengües. A més de la cerca de recursos textuals, també permet la cerca d'imatges o gràfics, de missatges electrònics a grups de discussió públics i la cerca per navegació sobre el conegut Open Directory Project, una iniciativa pública i oberta per indexar per categories els recursos textuals del WWW.

llista de paraules buides (*stop word list*). S'apropa, doncs, a la indexació per paraules clau, tot i que no duu a terme cap anàlisi semàntica, sinó que el sistema es limita a comparar llistes de cadenes de caràcters i excloure totes les que estiguin recollides a la llista de paraules buides. D'aquesta manera, s'alleugereix l'índex analític o índex invertit i es disminueix el temps de resposta del motor de cerca. A més, per tal de facilitar uns resultats més acurats i ajudar a l'usuari a decidir quin és el document adequat per a les seves necessitats, sovint, a més de l'adreça URL d'accés a cada document proposat, també es faciliten dades sobre el document que el motor de cerca ha recollit, com ara la data de creació o de recollida del registre a la base de dades, el títol i les primeres línies o un extracte del document.

- La formulació de consultes amb un motor de cerca

Tot i que cada motor té les seves peculiaritats, en general tots permeten expressar consultes complexes semblants amb condicionants de diferent naturalesa. En aquest àmbit, Google és un dels més complets ateses les prestacions següents:

- Cercar incloent-hi necessàriament una paraula (+) o excloent-la (-): simula la lògica Booleana i també permet incloure en la cerca paraules que el motor generalment ignora, com ara paraules buides.
- Cercar una frase o una unitat de més d'una paraula (entre cometes): és el que en algunes bases de dades es coneix com a cercar *as is* o coincidència exacta (ex. Eurodicautom). Si no es coneix una de les paraules que conformen la frase o se'n dubta, es pot substituir per *.
- Refinar una cerca ja realitzada afegint-hi altres paràmetres.
- Restringir la cerca per llengua, per freqüència mínima d'aparició d'una paraula o per domini o lloc web (site:).

- Cercar a la memòria caché del motor (caché:), on emmagatzema una còpia de cada cop que un document ha estat indexat o revisat; d'aquesta manera es pot accedir a documents que ja no es troben al web.[16]
- Cercar pàgines que tenen enllaços cap a una pàgina determinada (link:).
- Cercar pàgines similars a una pàgina determinada (related:).
- Cercar la informació que el motor té sobre una pàgina determinada (info:).
- Cercar corregint l'ortografia dels paràmetres de cerca (spell:).
- Cercar en pàgines relacionades amb l'àmbit de la borsa (stocks:).
- Cercar documents que continguin totes les paraules que se cerquen al seu títol (allintitle:).
- Cercar documents que continguin totes les paraules que se cerquen a la seva URL (allinurl:).
- Cercar documents que continguin alguna de les paraules que se cerquen a la seva URL (inurl:).

Altres motors, com ara AltaVista, permeten la cerca per truncació (*) i per proximitat (Near/n), de manera que recuperi documents en què una paraula *A* estigui a menys d'*n* paraules de distància d'una paraula B; una altra possibilitat és cercar en els continguts dels camps de metadades, que es troben a l'encapçalament del document (els més habituals són *Title*, *Author*, *Keywords* i *Description*).

Tots aquests criteris de cerca, amb els quals en cap cas no es realitzen cerques estrictament conceptuals o en funció del tema que aborden els documents, permeten la formulació de consultes amb un alt grau de complexitat emulant una consulta en llenguatge natural, i la consegüent obtenció de resultats amb un alt nivell de precisió (adequats a la consulta), tot i que es fa difícil establir la taxa de recuperació, atès que resulta impossible conèixer tots els documents rellevants existents al *World Wide Web*, o fins i tot els que formen part de la base de dades del motor.

[16] A més de l'índex de documents analitzats, Google també compta amb un segon índex de documents dels quals sap que existeixen (té la seva URL), tot i que encara no han estat indexats i que, per tant, no es troben a la seva memòria cau.

Actualment, molts motors de cerca, per tal de millorar el seu servei, també han apostat per bases de dades de grans dimensions, que cobreixin el màxim possible de llocs web. A finals de 2001, Google va fer públic que a la seva base de dades hi havia recollits més de 1.500 milions de documents, que actualitzava periòdicament (Sullivan, 2002); al juny de 2002 Fast[17] anuncià que la seva base de dades superava els 2.100 milions de pàgines indexades (Sherman, 2002). Tot i així, entre tots els motors i directoris no arriben a cobrir ni el 50% del que es calcula que és el *World Wide Web*, i que aproximadament el 10% de les seves dades estan desfasades o són errònies (Lawrence i Giles, 1999). En aquest prestigiós estudi, Lawrence i Giles també aconseguiren demostrar que menys del 4% dels registres de les bases de dades dels diferents motors coincideixen, ja que els robots de cada motor comencen la seva exploració partint d'un fons de pàgines diferent. Tot i així, també van comprovar que totes les bases de dades recollien més registres de pàgines dels EUA que de la resta del món, en un percentatge que no es corresponia amb la realitat del web. Malauradament, aquest estudi no ha estat actualitzat i no descriu necessàriament els motors amb bases de dades de més d'un miler de milions de registres, per la qual cosa no es pot afirmar que la situació es mantingui, tot i que experts en l'observació i anàlisi del funcionament dels cercadors afirmen que no ha variat substancialment.[18]

Tal com afirmen Baeza-Yates i Ribeiro-Net: "Considering how the Web is traversed, the index of a search engine can be thought of as analogous to the stars in an sky. What we see has never existed, as the light has travelled different distances to reach our eye." (1999: 382). Aquesta mateixa volatilitat en l'estat del web, en relació amb el qual la base de dades d'un motor de cerca, o d'un directori, sempre és una fotografia del seu passat, i no del seu present, és el que ens pot fer afirmar que una cerca mai no es pot dur a terme sobre la totalitat real i absoluta del web. A més, tal com veurem a

[17] Índex dels motors de cerca AllTheWeb.com (http://www.alltheweb.com) i Lycos (http://www.lycos.com).

[18] Afirmació que es desprèn dels comentaris realitzats per experts en articles recollits a Search Engine Watch (http://searchenginewth.com) o W3C, *World Wide Web* Consortium (http://www.w3.org), per exemple.

continuació, això no es deu únicament al fet que el web estigui en una situació de canvi continu, sinó que els robots dels motors no poden arribar a tots els seus racons.

c) El web invisible

La part del *World Wide Web* que no està recollit a les bases de dades dels cercadors rep el nom de **Web invisible** (*invisible Web*) [19]. Tot allò que els robots dels motors de cerca no veuen i, per tant, no indexen, resta invisible als ulls dels cercadors, i consegüentment també dels usuaris que els utilitzen com a porta d'accés a la informació.

Les bases de dades dels cercadors emmagatzemen llocs web que contenen **pàgines estàtiques**, generats i penjats d'un servidor pel seu autor o responsable, el contingut dels quals es manté estable fins que aquest autor o webmaster decideix canviar-ne el contingut afegint o eliminant pàgines o modificant el contingut de les existents. Els motors de cerca hi arriben gràcies a enllaços que surten d'altres llocs web, l'indexen, i periòdicament el revisen per tal d'actualitzar i incloure a la base de dades del motor de cerca les modificacions que s'hi puguin haver realitzat.

Gran part de la informació que es pot consultar al WWW no es troba en pàgines estàtiques, sinó en pàgines que es generen després de fer una consulta en línia sobre una base de dades. Es tracta, doncs, de **pàgines dinàmiques**, que un sistema genera automàticament i sovint només de manera virtual, ja que no queda indefinidament al web, sinó que només es podrà tornar a *visitar* si es torna a fer la consulta que la va originar. És el cas, per exemple, de les pàgines resultants de consultar una base de dades terminològica com Eurodicautom, o d'una guia telefònica, o fins i tot de realitzar una cerca sobre un motor de cerca.

[19] Sovint rep altres denominacions equivalents, com ara web profund (*deep Web*) o Internet amagat (*hidden Internet*) (Tracey, 2002). La denominació web profund respon a la metàfora dels tipus de pesca: existeix el web superficial, al qual és fàcil accedir, i el web profund, de més difícil accés.

El conjunt d'informació que es pot recuperar mitjançant pàgines dinàmiques és el major component del web invisible, tot i que també n'hi ha d'altres tipus. El fet que un robot no pugui accedir a la informació d'un lloc pot obeir a qualsevol dels motius següents:

- Causes tècniques:
  - Part del lloc, o tot ell, està format per pàgines dinàmiques. Tot i que no es pot indexar tota la informació del lloc web, els motors sovint permeten indexar les seves planes estàtiques, encara que siguin poques.
  - Per accedir al lloc es requereix una identificació. Són llocs restringits (intranets o llocs privats) o d'accés públic, però que demanen al visitant que s'identifiqui.
  - Les pàgines estan en formats que els robots no poden llegir. Tots els robots llegeixen documents en format HTML, però cada cop és més habitual trobar informació al WWW en altres formats (com ara PDF, formats multimèdia o formats propietaris).[20]
  - Pàgines generades amb el llenguatge de programació *script*, que habitualment inclouen a la seva URL un signe d'interrogació (?).[21]
  - Llocs que no són destí de cap enllaç des d'altres documents, per exemple perquè són nous o perquè l'autor no l'ha promocionat.
- Causes polítiques:
  - Informació que els cercadors no volen indexar. Es tracta de pàgines amb continguts que els cercadors consideren incorrectes o prohibits i als quals no volen facilitar l'accés, com per exemple continguts xenòfobs o pornogràfics. Els cercadors exclouen aquesta mena de pàgines per decisió pròpia o a causa de la legislació dels diferents països.
  - Informació que els servidors en què es troben no permeten que sigui indexada. Els servidors poden vedar l'accés als robots per preservar la

[20] Excepcions: Google pot indexar i recuperar pàgines en format PDF, i també pàgines formades per imatges, tal com també fa AltaVista.

[21] Els articles de la majoria dels diaris en línia són un exemple de pàgines generades amb *script*. Tot i que els robots sí que poden llegir aquesta mena de pàgines, no ho fan perquè poden trobar-hi trampes per a robots que els faria caure amb bucle infinit del qual no podrien sortir.

informació que contenen fora del seu abast o per no saturar-se amb les visites dels robots.

Un estudi realitzat per l'empresa BrightPlanet (Bergman, 2001), revela que el web profund, amb aproximadament 7.500 terabytes d'informació, és cinc-cents cops més voluminos que el web superficial i té molta més qualitat, i està format per aproximadament 550.000 milions de documents diferents dels quals aproximadament 200.000 existeixen amb adreça URL pròpia.

El web invisible conté recursos de diferent naturalesa. Els principals són aquests:

- Bases de dades que es poden consultar en línia, sobretot bases de dades especialitzades en un tipus de dades concret o en un àmbit determinat (bases de dades de publicacions mèdiques, de patents o guies de telèfon, per exemple).
- Eines interactives: tota mena de calculadores, conversors, etc.
- Material de consulta propi d'arxiu, com ara, documentació d'institucions públiques.
- Fonts estadístiques, públiques o privades, amb dades socials o geogràfiques, per exemple.
- Llocs de notícies: agències de premsa, revistes, diaris, cadenes de televisió, etc., que generen informació per ser publicada a la web.
- Portals: llocs compostos per elements de diferent naturalesa (notícies, grups de discussió o recursos interactius, per exemple) dedicats a un sector social o un àmbit temàtic concret.
- Catàlegs de biblioteques i serveis d'informació.
- Intranets: llocs restringits als membres d'una empresa o una institució per facilitar l'intercanvi d'informació interna.
- Obres de consulta (com ara diccionaris en línia).

Les bases de dades són el principal component del web invisible, puix que s'estan convertint en un dels mecanismes més útils per accedir a informació altament estructurada. A més, d'aquesta manera se superen les limitacions de l'accés a documents mitjançant sistemes de recuperació basats en índexs o llistats, que no

representen una resposta suficientment vàlida davant el creixement exponencial dels recursos d'informació (López Medina, 2001: 8). La consulta d'aquestes bases de dades es realitza mitjançant un formulari o una passarel·la que transmet la pregunta que l'usuari realitza al web que conté la base de dades i permet mostrar el resultat de la consulta en una pàgina dinàmica.

Segons l'estudi de BrightPlanet, les bases de dades especialitzades, dedicades a un tema específic, agrupen més de la meitat de la informació englobada al web invisible (vegeu figura 3-2, pàgina 159), i més del noranta-cinc per cent del total del web invisible és d'accés públic. La major part de la informació inclosa a bases de dades i altres recursos pertanyents al web invisible ha estat validada o revisada per experts o s'ha elaborat seguint un procés escrupulós que en garanteix la qualitat. El lloc prototípic del web invisible és molt més voluminós que els llocs del web superficial, ja que el primer té una mitjana de 5,43 milions de registres, tot i que la mitjana global n'és 4.950. Aquests llocs reben el doble del trànsit i són el destí de gairebé el doble d'enllaços que un lloc web comú; aquestes dades confirmen que el contingut informatiu del web invisible és d'alta qualitat.

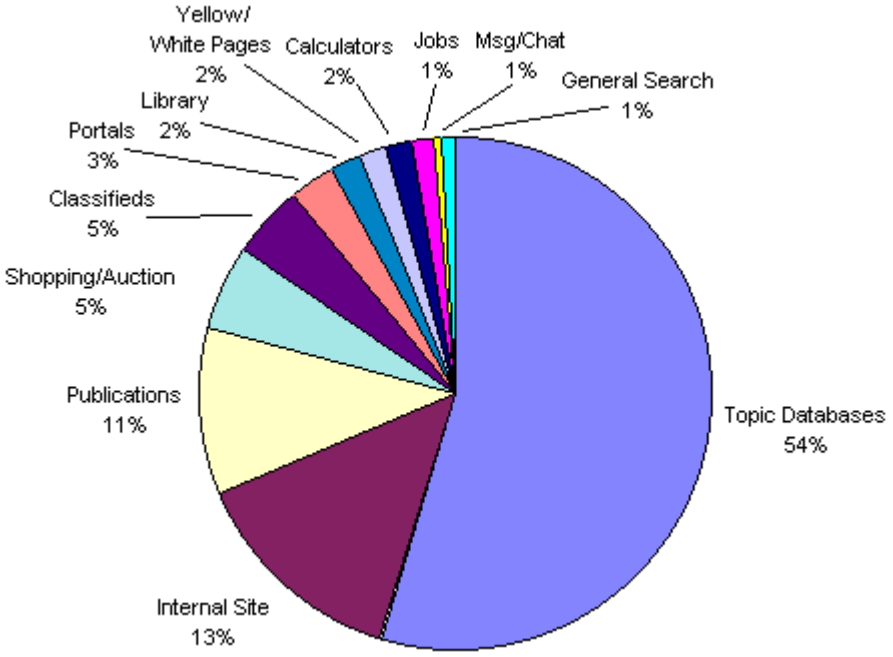

Figura 3-2. Gràfic representatiu dels recursos que es troben al web invisible segons l'estudi realitzat per BrightPlanet (Bergman, 2001)

Els diferents continguts temàtics recollits principalment a les bases de dades especialitzades i a les publicacions, que sovint també estan dedicades a un tema específic, també han estat objecte d'estudi per part de Bright Planet, que van arribar a resumir-los en categories que no sempre es veuen reflectides als directoris, i que tampoc no responen a les tradicionals classificacions del coneixement humà:

Agricultura	2,7%	Humanitats	13,5%
Belles arts	6,6%	Informàtica / Web	6,9%
Ciència, matemàtiques	4%	Negocis (economia)	5,9%
Compres	3,2%	Notícies, mitjans de premsa	12,2%
Dret/Política	3,9%	Oci, esport	3,5%
Educació	4,3%	Referències	4,5%
Enginyeria	4%	Salut	5,5%
Estil de vida	4%	Treball (ocupació)	4,1%
Gent, empreses	4,9%	Viatges	3,4%
Govern	3,9%		

Taula 3-5. Abast temàtic del web invisible segons l'estudi de BrightPlanet (Bergman, 2001)

Per tal de facilitar l'accés als recursos del web invisible, han aparegut cercadors que els recullen ordenant-los en categories. L'usuari cerca recursos per temes o paraules clau, i el cercador respon proposant-li llocs web on podrà satisfer la seva necessitat informativa. Alguns, com Complete Planet[22], fins i tot faciliten la consulta d'algunes de les bases de dades mitjançant un formulari des del mateix cercador.

[22] Complete Planet: http://www.completeplanet.com/. És el cercador de l'empresa BrightPlanet i s'accedeix a la totalitat dels seus serveis per subscripció.

De tota manera, davant una necessitat informativa, qualsevol internauta recorre en primer terme a un cercador genèric; aquests sistemes d'RI, que no poden consultar directament les dades que es troben al web invisible atès que és necessari un cert grau d'interacció (emplenar un formulari de consulta o identificar-se com a usuari), no poden extreure-hi dades. Tanmateix, el que sí que fan es facilitar l'accés als llocs del web invisible, ja que acostumen a indexar les planes estàtiques d'aquests recursos (sovint la seva pàgina inicial o una descripció del recurs). Aquest és un dels motius pels quals el límit entre el web visible i l'invisible no es pot establir amb claredat, sinó que és una zona de clarobscurs, la denominada zona grisa o **web soma** (*shallow Web*)[23]. Aquest subconjunt del web està format per recursos de diferent naturalesa:

- Recursos propis del web superficial que, per no estar indexats (a causa del seu format o d'una manca d'enllaços dirigits cap a ell), formen part del web desconegut.
- Recursos dels quals no es coneix el responsable secundari o canvien d'URL amb assiduïtat. Aquests llocs no els recull cap índex de recursos del web profund en dubtar de la qualitat del seu contingut.
- Recursos que formalment s'inclouen al web superficial, tot i que el seu contingut prové parcialment o en la seva totalitat del web profund (pàgines estàtiques que han inclòs el contingut d'altres pàgines dinàmiques, o bases de dades que ofereixen la seva informació en pàgines estàtiques i dinàmiques alhora).

L'obtenció d'informació de manera automàtica o semiautomàtica del web invisible ha rebut la denominació de **mineria de dades**[24] (*data mining*) (Clyde, 2002) i, actualment, com pràcticament tot al *World Wide Web*, és una línia de recerca oberta

[23] Denominació anglesa encunyada per Danny Sullivan (2001) seguint la metàfora de la pesca. Seguint aquesta metàfora, proposem la denominació Web soma, tot posant de relleu el seu caràcter de poca profunditat.

[24] Calc de la denominació anglesa molt estès, recollit, per exemple, per Sangüesa Solé i Molina Félix al seu curs de formació continuada titulat *Data mining*, a la Universitat Oberta de Catalunya.

que té com a objectiu la generació de sistemes d'RI que permetin cercar informació directament sobre les fonts del web invisible, augmentant la qualitat de la informació recollida a la resposta. Ara com ara, la mineria de dades és un servei que ofereixen els professionals de la informació, ja que són els que tenen els coneixements i les habilitats necessàries per fer-ho, amb l'aixopluc de serveis d'informació, que són els que compten amb el programari que sovint és necessari per consultar determinats recursos del web invisible, així com les claus d'accés.

d) Altres sistemes de cerca

Entre els cercadors, a més dels directoris i els motors de cerca, també hi ha els **metacercadors**[25], sistemes d'RI que no cerquen sobre una base de dades pròpia, sinó que transmeten la consulta a un conjunt de directoris i motors de cerca. Són sistemes capaços d'enviar consultes a diversos cercadors alhora i proporcionar al seu usuari una única llista integrada amb tots els resultats obtinguts. Els components habituals d'un metacercador són els següents:

- Selector de cercadors: sistema mitjançant el qual l'usuari pot determinar sobre quins cercadors vol que es dugui a terme la seva consulta.
- Agents d'interfície: tradueixen la consulta al format de cada cercador o motor de cerca sobre els quals es farà la consulta.
- Visualitzador dels resultats: sistema que recull les dades obtingudes de les consultes realitzades sobre els cercadors escollits. Sovint n'elimina els resultats duplicats i els ordena en funció de la seva rellevància. (Repman i Carlson, 1999).

Els metacercadors permeten expressar consultes utilitzant operadors Booleans, o cercar frases exactes, però habitualment no suporten cerques amb paràmetres molt específics, com ara els que utilitza Google. A més, en recollir els resultats obtinguts

[25] Per exemple, MetaCrawler: http://metacrawler.com

amb cada directori o motor de cerca, també acostumen a limitar el nombre de resultats provinents de cada cercador que inclourà la llista de resultats final.

Al *World Wide Web* també es troben llocs web que s'anuncien com a metacercadors, tot i que no permeten realitzar cap cerca conjunta: en realitat són llistats de directoris i motors de cerca que permeten cercar únicament sobre un sol dels recursos que proposen.

Una altra manera de consultar diversos cercadors alhora és recórrer a **agents de cerca intel·ligents**[26] (*intelligent agents*) (Gómez, 1999; Oppenheim *et al.*, 2000: 193; Vilarnau, 2001: 44), aplicacions informàtiques que s'instal·len a l'estació de treball de l'usuari, no es consulten via web com els cercadors. Aquests sistemes, amb totes les característiques dels metacercadors, acostumen a oferir altres elements addicionals, com ara:

- Desar el resultat de la cerca al disc dur.
- Refinar la cerca sobre els resultats obtinguts.
- Validar els resultats, eliminant, per exemple, tots aquells resultats que en realitat són enllaços morts perquè la pàgina final ja no es troba al web.
- Descarregar les pàgines que han estat proposades en els resultats de la cerca.
- Programar tasques: permeten, per exemple, repetir cerques periòdicament de manera automàtica.

El món dels sistemes d'RI està en continu procés de recerca, en part per les mancances detectades amb el *descobriment* del web invisible, i en part perquè els diferents sistemes encara no han arribat al nivell de qualitat desitjat. Amb l'aparició de motors de cerca com, per exemple, Teoma, Vivisimo i Wisenut[27], s'aposta per un sistema d'indexació de la informació que permeti la generació automàtica de categories

[26] Per exemple, Copernic (en les seves diferents versions): http://www.copernic.com.

[27] Teoma: http://www.teoma.com; Vivisimo: http://vivisimo.com; Wisenut: http://www.wisenut.com.

temàtiques en funció dels criteris expressats per l'usuari en la seva consulta, configurant així un híbrid entre el directori i el motor de cerca prototípics, tant en la fase d'indexació con en la de recuperació de la informació.

Segons els experts (Baeza-Yates i Ribeiro-Net, 1999; Millán, 2000), en el futur els cercadors no cercaran meres cadenes de caràcters, sinó que s'introduiran mecanismes de suport morfològic (que en cercar, per exemple, un infinitiu recuperaran documents on aparegui qualsevol forma conjugada d'aquell verb) i de suport semàntic (que permetrà la recuperació de documents que continguin expressions sinònimes a les que s'inclouen en la cerca). L'objectiu final sembla que és, doncs, refinar els mecanismes d'indexació i recuperació de la informació per permetre l'expressió de consultes de la manera més propera al llenguatge natural possible, i fins i tot sobrepassar els límits d'una llengua per arribar a sistemes multilingüístics, on es pugui expressar una consulta en una llengua i obtenir resultats rellevants en qualsevol altra llengua. Malauradament, però, no sembla aquest un objectiu que es pugui assolir satisfactòriament a curt termini.

Això no obstant, una altra manera de millorar el rendiment dels diferents sistemes d'RI en l'entorn web passa per la implementació de mecanismes d'indexació automàtica que no estiguin basats en el contingut textual dels recursos web, com fins ara, sinó en una descripció realitzada amb aquest fi per l'autor o responsable de cada recurs. Aquesta línia d'actuació es basa en l'explotació del potencial dels camps de metadata de tot document en format HTML (o variacions) amb la finalitat de facilitat la catalogació dels recursos, o dit d'una altra manera, que cada recurs vagi acompanyat per una mena de *fitxa descriptora* per tal que la seva indexació en les bases de dades dels diferents sistemes de cerca del web resulti molt més fiable i permeti així una extracció d'informació molt més precisa.

La iniciativa més coneguda en aquest sentit és la norma *Dublin Core*[28] (Weibel, 1998), que advoca per la incorporació de la següent sèrie de camps en l'encapçalament dels documents en format HTML: títol, autor, àrea temàtica (indicada amb paraules clau, seguint un vocabulari controlat i fins i tot un sistema de classificació del coneixement), descripció, responsable secundari, contribucions (a més de l'autor), data, tipus de recurs (en funció d'una categorització de recursos web pròpia), format, adreça URL, font (en el cas que s'inclogui informació que prové d'altres llocs), llengua, relació amb altres recursos, abast de la informació (tant en l'espai, regional o global, com d'altra mena) i indicació de drets de diferent tipus (d'autor, copyright, etc.).

La norma *Dublin Core* ha estat adoptada pel *World Wide Web* Consortium (W3C)[29], responsable del desenvolupament de tecnologies que permetin que el *World Wide Web* assoleixi el seu màxim potencial com a entorn d'informació, comerç i comunicació, per la qual cosa és molt possible que s'acabi aplicant de manera sistemàtica, almenys en àmbits acadèmics i en els relacionats amb serveis d'informació.

Per tots els motius exposats al llarg d'aquest apartat, des de la distribució de la informació al *World Wide Web* fins al funcionament dels cercadors i la part del web que queda fora del seu abast, es pot arribar a la conclusió que, tot i que molta informació queda fora de les bases de dades dels cercadors, també és molta la informació que abracen. No obstant això, el gran inconvenient que troba el traductor usuari dels cercadors per accedir a les diferents fonts d'informació en línia és, un cop troba els recursos que s'adeqüen temàticament a les seves necessitats, la impossibilitat de destriar els recursos de qualitat dels que no ho són. Probablement, si finalment es generalitza l'ús de la norma *Dublin Core*, o cap altre sistema similar, el traductor podrà prendre decisions de caire documental en funció de nous criteris que, sense analitzar el contingut dels documents, li puguin garantir la seva qualitat, com ara la procedència del recurs o el seu autor.

---

[28] Dublin Core Metaata Initiative: http://dublincore.org. La *Dublin Core Metadata Initiative* es materialitzà en un seminari organitzat conjuntament pels organismes de recerca NCSA i OCLC a la ciutat de Dublin, Ohio, el març de 1995.

[29] W3C: http://www.w3.org.

2. La cerca de documents digitals especialitzats

La cerca de documentació digital es duu a terme en funció d'un propòsit concret. No se cerca informació en abstracte i fora de context, sinó que es fa amb un objectiu determinat. És per això que, a més d'observar les característiques de la informació tal com s'ofereix a Internet, a l'hora de realitzar una cerca també cal tenir en compte les necessitats informatives del que serà el seu receptor. En aquest sentit, Chowdhury, en reflexionar sobre la **necessitat d'informació** (*information need*) arriba a les conclusions següents:

> "1. information need is a relative concept. It depends on several factors and does not remain constant; 2. information needs change over a period of time; 3. information needs vary from person to person, from job to job, subject to subject, organization to organization, and so on; 4. people's information needs are largely dependent on the environment. For example, information needs of those in an academic environment are different from those in an industrial, business or government/administrative environment; 5. measuring (quantifying) information need is difficult; 6. information need often remains unexpressed or poorly expressed; 7. information need often changes upon receipt of some information." (Chowdhury, 1999: 181)

Durant el procés de documentació del traductor especialitzat, el concepte de necessitat d'informació també és relatiu, no només perquè els traductors tenen necessitats informatives diferents de les d'altres professions, sinó sobretot perquè les necessitats de cada traductor varien en funció del seu encàrrec de traducció. Si bé per informació de qualitat s'entén tota informació objectivament completa i actual, la traducció que està duent a terme pot obligar al traductor a cercar informació que no coincideix plenament amb aquestes característiques. Així doncs, en traduir un text científic anterior, per exemple, a un descobriment rellevant que modifica els fonaments teòrics d'aquell domini, el traductor s'haurà de documentar amb textos coetanis amb el seu original, o que tractin el tema amb la mateixa perspectiva. En altres casos, si en analitzar l'original el traductor estableix que es tracta d'un text propi, per exemple, d'una situació comunicativa concreta o amb uns condicionants pragmàtics determinats, on no s'empra el to típic de la comunicació entre experts, probablement, a causa de la

distància que s'estableix entre els interlocutors, cercarà textos produïts o utilitzats en condicions similars per documentar-se.

Per tot això, i coincidint en certa manera amb la sisena conclusió de Chowdhury, el traductor ha de conèixer molt bé les seves necessitats d'informació en cada encàrrec i ha de poder-les expressar de la manera més acurada possible per tal d'obtenir els documents més adients en cada cas. A més, el protagonisme de la informació en la tasca del traductor cobra una especial rellevància, tal com assenyala Mayoral:

> "Creemos que lo que caracteriza al traductor frente a otros especialistas, como químicos o ingenieros, es su misión como comunicador de saber especializado entre lenguas diferentes: el traductor no es el especialista que produce o recibe información especializada sino el profesional cuya principal responsabilidad consiste en difundir esta información salvando barreras lingüísticas y culturales." (Mayoral, 1997/1998: 137)

D'aquesta manera, l'afirmació de Mayoral ratifica l'obligació que té el traductor de conèixer les seves necessitats informatives en profunditat i com solventar-les. En aquesta tesi ens centrem en la cerca i obtenció d'informació exclusivament a partir de recursos textuals digitals extrets del web per fer-los servir com a textos paral·lels (vegeu l'apartat 1 "La documentació amb el text paral·lel en traducció" del capítol 5, pàgina 239).

2.1. La cerca dels recursos textuals digitals especialitzats sobre els Leònids per a la posterior creació de corpus monolingües comparables

A l'hora de cercar recursos textuals digitals en el marc d'aquesta recerca, i atès que després aquests recursos seran la base d'un estudi lingüísticoconceptual partint de la metodologia de la lingüística de corpus, s'ha volgut prioritzar el factor quantitatiu per sobre del qualitatiu. Per aquest motiu, hem dut a terme cerques amb una taxa de recuperació alta (vegeu l'apartat 1.2.2 "La selecció i obtenció d'informació" d'aquest capítol, pàgina 140) i hem deixat en un segon pla la taxa de precisió.

Tal com s'indicava a la introducció, hem escollit com a tema científic sobre el qual treballar el fenomen astronòmic dels Leònids. Aquesta elecció respon fonamentalment a la nostra voluntat de comptar amb un corpus de textos finals que reflectís un ventall complet de textos especialitzats possibles, ja que:

- Aquest fenomen es pot observar amb relativa facilitat, sense necessitat d'accedir a tecnologia molt desenvolupada. Per aquest fet, prevèiem que fou objecte d'estudi d'experts i semiexperts, i que tots dos col·lectius publicarien reflexions o experiències al respecte a la *World Wide Web*.
- Els astrònoms representen un col·lectiu avesat en la tecnologia en l'àmbit d'usuari, per la qual cosa esperàvem trobar textos a Internet publicats per tota mena d'usuaris: especialistes i semexperts (aficionats, periodistes, etc.).

El procés de cerca de recursos textuals sobre Leònids, el domini en què se centra aquesta investigació, s'ha realitzat amb l'ajuda de l'agent intel·ligent de cerca Copernic[30]. De les diferents opcions que ofereix aquest programa s'han fet servir les següents:

1. Cerca simultània sobre els directoris i motors de cerca següents: Altavista, AOL.com search, CompuServe, Direct Hit, Euroseek, Excite, FAST Search, Google, HotBot, LookSmart, Lycos, MSN Web Search, NBCi, Netscape Netcenter, Open Directory Project i Yahoo.
2. Limitació dels resultats de la cerca a 1.000 per cercador.
3. Validació dels resultats obtinguts, eliminant-ne els enllaços morts.

[30] Durant el procés de documentació, el programa ha canviat en dues ocasions de versió, per la qual cosa s'han utilitzat les versions gratuïtes del programa següents: Copernic 2000, Copernic 2001 i Copernic Agent Basic, com també la versió comercial Copernic 2001 Pro.

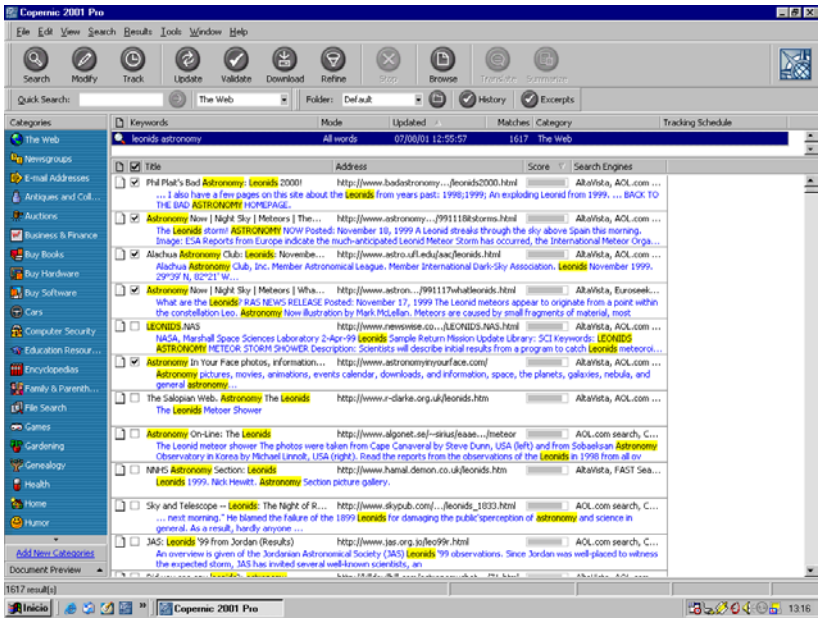

Figura 3-3. Interfície de Copernic 2001 Pro (e. p.)

Les cerques s'han fet en tots els casos amb dues paraules clau: el domini i l'aspecte específic sobre el qual haurà de tractar el corpus, és a dir, Leònids. Ateses les diferents possibilitats de redacció en cadascuna de les llengües (anglès, castellà i català), incloent-hi grafies errònies, les combinacions utilitzades per cercar documents han estat les següents:

Llengües	Combinacions de cerca	
Anglès	- "leonids" + "astronomy" - "leonid" + "astronomy"	
Castellà	- "leónidas" + "astronomía" - "leónida + "astronomía" - * "leonidas" + "astronomia" - * "leonida" + "astronomia"	
Català	- "leònid" + "astronomia" - "leònids" + "astronomia" - * "leonides" + "astronomia" - * "leonida" + "astronomia" - * "leonid" + "astronomia" - * "leònides" + "astronomia" - * "leònida" + "astronomia"	- * "lleònides" + "astronomia" - * "lleònida" + "astronomia" - * "lleònid" + "astronomia" - * "lleonides" + "astronomia" - * "lleonida" + "astronomia" - * "lleonid" + "astronomia"
Català 2	- "pluja de meteorits" - "pluges de meteorits" - "pluja de meteors" - "pluges de meteors"	- "pluja d'estels" + astronomia - "pluges d'estels" + astronomia
Català 3	- "meteorit" - "meteorits"	- "meteor" - "meteors"

Taula 3-6. Combinacions de cerques en anglès, castellà i català (e. p.)

Com es pot comprovar a la taula 3-6, en totes tres llengües s'ha previst la mateixa combinació en singular i plural. En castellà, a més, també s'ha previst la grafia sense accent (*leonidas, *leonida) que, tot i ser incorrecta, es pot trobar sovint als documents d'Internet, ja sigui per un error ortogràfic o per la impossibilitat d'utilitzar accents amb teclats o sistemes operatius anglosaxons, o almenys la impossibilitat de fer-ho amb comoditat.

La cerca en català va resultar més complexa que en les altres llengües a causa fonamentalment dels motius següents:

a) La manca de consens ortogràfic. Malauradament, el terme "Leònids" té una denominació normalitzada, per organismes terminològic i filològic que els experts segueixen en poques ocasions. És per això que, en adoptar la forma normalitzada Leònids, deixaríem de banda forems com "leònides", que és la

que es recull en els recursos textuals publicats al document del Departament de Meteorologia i Astronomia de la Universitat de Barcelona[31].

b) A més d'aquesta inconsistència denominativa, també han estat tingudes en compte combinacions ortogràficament incorrectes, marcades amb un asterisc.

c) Atès el baix nombre de recursos textuals obtinguts a partir de les combinacions amb "Leònids" o qualsevol altra grafia que representi aquest concepte, es va decidir ampliar la cerca utilitzant fórmules hiperonímiques, com són ara "pluja d'estels"[32], "pluja de meteorits", "pluja de meteors", i posteriorment també amb "meteorit" i "meteor" en totes les seves possibles grafies. Aquesta decisió es pren després d'establir la relació d'hiperonímia entre aquestes unitats mitjançant una breu anàlisi del corpus català original (vegeu el capítol 6 "L'anàlisi del corpus compilat", pàgina 283).

Les cerques de documents s'han repetit cíclicament durant tres anys (del 1999 al 2001) ja que, com es podrà comprovar amb l'anàlisi dels textos, el fenomen astronòmic denominat Leònids es produeix un cop cada any, a mitjan novembre, per la qual cosa cada any es generen nous recursos textuals digitals després de l'observació del fenomen. Molts dels recursos obtinguts en les successives cerques ja havien estat identificats prèviament i es descartaven per tal d'evitar repeticions innecessàries. L'última recollida de dades es realitzà el desembre de l'any 2001.

Tal com s'ha assenyalat al començament d'aquest apartat, aquests criteris de cerca permeten realitzar cerques amb una taxa de recuperació alta. A causa de la senzillesa de la consulta, aquesta es pot realitzar sobre tots els cercadors sense haver-ne de modificar la sintaxi. Tot i que d'aquesta manera s'aconsegueix el major nombre de resultats possible, també és cert que el soroll pel que fa a l'heterogeneïtat del contingut i la naturalesa dels recursos digitals obtinguts resulta molt alt. És per això que es

[31] Per exemple, el document CA00301 del corpus monolingüe català.

[32] En aquest cas es va decidir cercar "pluja d'estels" + "astronomia" pel gran ús que es fa d'aquesta expressió com a imatge en textos literaris.

revisaren totes les propostes obtingudes a partir de les consultes realitzades amb Copernic i alguns es descartaren en funció dels criteris següents:

- Contenir la paraula clau ("Leònids") únicament a la llista d'enllaços del recurs, però no al cos del text.
- A causa de la polisèmia del mot Leónidas (castellà) i Leonid (anglès), que a més de referir-se al fenomen astronòmic també poden ser un nom propi, algunes planes foren descartades perquè no estaven dedicades a aquest fenomen, tot i que contenien les dues paraules clau: Leònids i astronomia.
- Sovint, un mateix document hipertextual apareix en més d'una ocasió sota adreces URL diferents, per la qual cosa se n'han descartat els duplicats.
- Contenir text en més d'un idioma, per raó de la impossibilitat d'analitzar-lo apropiadament.
- Tractar-se en realitat d'un missatge de correu electrònic, ja que considerem que aquest mitjà de comunicació està a cavall entre la comunicació escrita i l'oral. Els missatges de correu electrònic poden considerar-se en certa mesura com a part d'una conversa, ja que cada missatge és una intervenció en un diàleg de dues o més persones. Per aquest motiu, en un missatge es pot fer referència a una unitat d'informació que ha aparegut prèviament a la conversa sense citar-la explícitament (Calsamiglia i Tusón, 1999: 33); altrament, sovint es transgredeixen les normes habituals de redacció tot utilitzant icones o abreviatures pròpies d'aquest medi i que exigirien una anàlisi específica.

Per tot això, de tots els recursos textuals obtinguts mitjançant les diferents cerques es va tenir en compte finalment un nombre relativament baix. La major part dels recursos descartats ho van ser perquè "Leònids" (en cadascuna de les llengües) apareixia únicament en un enllaç, i no al cos del recurs.

	Anglès		Castellà		Català Leònids		Català Hiperònims de Leònids	
Adreces de recursos **vàlids**	406	27,45%	39	3,72%	14	3,57%	22	11,46%
Adreces de recursos **no trobats**	230	15,55%	323	30,79%	64	16,33	21	10,94%
Adreces de recursos **no vàlids**	842	56,93%	687	65,69%	314	80,10%	149	77,60%
Adreces de recursos d'accés restringit[33]	1	0,07%						
Total	1.479		1.049		392		192	

Taula 3-7. Resultat de les cerques per idiomes (e. p.)

De tota manera, resulta significatiu que el percentatge de recursos vàlids en anglès (27,45%) sigui molt més elevat que els vàlids en català (3,57% i 11,46%) i castellà (3,72%). Creiem que aquest fet pot ser conseqüència de diferents factors, tot i que no s'ha pogut comprovar:

- Gran part dels recursos en castellà i català remeten a recursos en anglès, que són els que contenen la informació. Per aquest motiu, la relació entre recursos no vàlids i recursos vàlids, que en anglès es manté 3:1 aproximadament, en castellà i català és veu minvada (aproximadament 30:1 o 10:1[34]), perquè la informació no la trobem en aquestes llengües sinó en anglès.
- Atès que l'única limitació a l'hora de fer les cerques era de no recollir més de 1.000 resultats per cercador, es pot arribar a suposar que en castellà i català s'ha obtingut gairebé tots els recursos que coincidien amb els paràmetres de la cerca, fins i tot els més peregrins, mentre que en anglès s'han pogut recuperar únicament els recursos de major qualitat, deixant de banda els menys apropiats.

[33] Aquest recurs pertany al que s'ha anomenat web invisible, ja que per accedir-hi es necessita una clau d'identificació. Generalment, el contingut dels recursos textuals del web invisible no pot ser analitzat ni indexat pels cercadors.

[34] En el cas de la cerca en català dels hiperònims de "Leònids" pot resultar poc informatiu en aquest àmbit, puix que no s'ha realitzat una cerca equivalent ni en anglès ni en castellà.

Amb tot això, s'han identificat tota una sèrie de llocs web[35] que contenen recursos textuals digitals sobre el fenomen astronòmic dels Leònids. En xifres absolutes, aquests han estat els resultats finals de la cerca:

Llengües	Llocs Web identificats
Anglès	242
Castellà	26
Català	16[36]
Total	287

Taula 3-8. Llocs Web identificats en cada llengua que formaran part dels corpus monolingües (e. p.)[37]

[35] Entenem lloc web com el conjunt de pàgines web (recursos textuals amb informació en qualsevol altra morfologia) que es troben dins del mateix domini (http://...../). Un lloc web no és necessàriament un document hipertextual; de fet hi pot haver més d'un document dins del mateix lloc. Sovint resulta molt difícil establir els límits entre els diferents documents. Es podria dir, doncs, que lloc web és una unitat física d'arxius, mentre que document hipertextual és una unitat lògica.

[36] Deu llocs identificats a partir de les diferents combinacions de Leònids + astronomia; sis llocs identificats a partir de cerques realitzades amb els hiperònims de Leònids.

[37] La relació exacta dels recursos digitals inclosos als corpus figura a l'annex C.

3. L'extracció de documents digitals de la xarxa

Un cop identificats els llocs web que contenen els recursos textuals digitals que formaran part de cadascun dels corpus, cal descarregar-los de la xarxa per tal de poder-los analitzar posteriorment. Aquesta operació s'ha dut a terme amb l'ajuda d'un copiador de webs, en concret el programa Offline Explorer, de Metaproducts. Amb aquest programa copiem els documents hipertextuals sencers; és a dir, no descarreguem únicament els recursos textuals identificats en la fase anterior, sinó que descarreguem també el document al qual pertanyen.

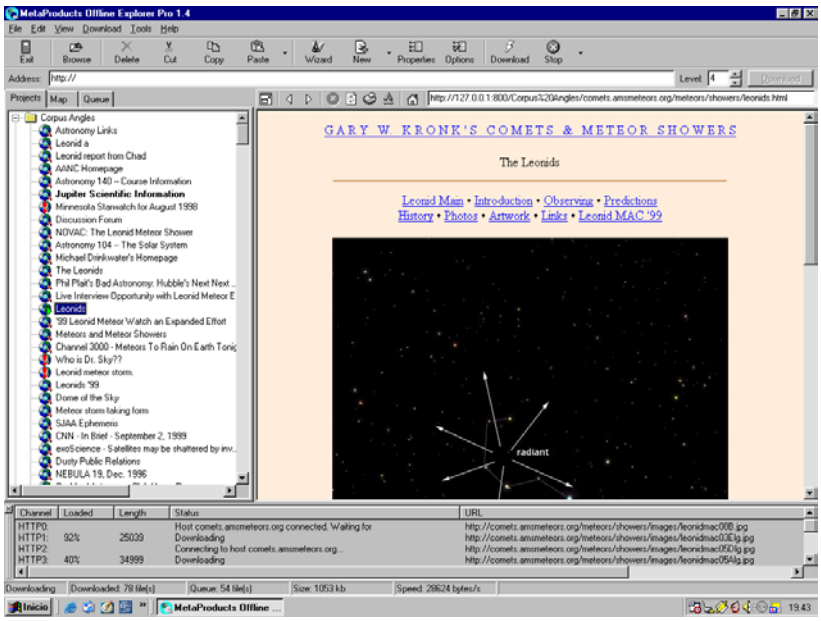

Figura 3-4. Interfície del copiador de webs Offline Explorer (e. p.)

D'aquesta manera, a més de l'arxiu (generalment en format HTML) que conté el recurs textual digital, també descarreguem qualsevol arxiu en un altre format que el complementa (imatges, vídeos, aplicacions interactives, etc.), així com la resta de recursos textuals que l'envolten en l'estructura del document hipertextual. A l'hora d'utilitzar el copiador de webs únicament s'han tingut en compte dues restriccions:

1. Descarregar arxius que pertanyen al mateix document hipertextual; en altres paraules, el programa només ha copiat el destí dels enllaços interns, però no dels externs.
2. No aprofundir més de quatre nivells; és a dir, el programa no ha descarregat arxius que es troben a més de quatre enllaços hipertextuals de distància de l'arxiu de partida, encara que es trobin dins del mateix document hipertextual.

Molts agents intel·ligents de cerca, com el programa Copernic utilitzat en la fase d'identificació dels recursos textuals, també compten amb una funció que permet descarregar planes d'Internet. Aquests programes poden descarregar els arxius proposats en la llista de resultats d'una cerca, però no els documents als quals pertanyen, per la qual cosa ens vam decantar per la utilització del copiador de webs.

En descarregar tant cadascun dels recursos identificats com els que els envolten augmentem el volum d'informació susceptible de ser analitzada com a font d'informació per a la traducció creix. De tota manera, i per tal d'evitar un soroll innecessari provocat per recursos textuals provinents de llocs heterogenis i que no tenen cap relació amb l'àmbit temàtic sobre el qual estem treballant, descartarem tots aquells recursos que no continguin la paraula "Leònids" (en cadascun dels idiomes i en les diferents grafies previstes).

Llengües	Volum d'informació descarregada
Anglès	6.522 mb[38]
Castellà	409 mb
Català	190 mb
Total	**7.121 mb**

Taula 3-9. Volum total d'informació descarregada d'Internet (e. p.)

[38] Tot aquest volum d'informació es va poder recollir en onze CD-ROM.

Com a resultat d'aquest procés vam obtenir un gran volum d'informació, tal com mostra la taula 3-9, en la qual també es fa palès el gran desequilibri que hi ha entre el volum d'informació recollit en anglès i el recollit en les altres dues llengües.

3.1. El procés d'identificació dels recursos textuals dels corpus

Fins ara hem descrit la metodologia que hem seguit, en primer lloc, per identificar pàgines web a Internet que coincidissin amb la necessitat d'informació que volem cobrir i, en segon lloc, per descarregar els llocs web a què pertanyen per tal de tenir una versió el més completa possible del document hipertextual. A continuació vam identificar entre totes les pàgines descarregades aquells recursos textuals que acomplien els condicionants establerts per tal de formar part del corpus, tornant a aplicar els requisits indicats a l'apartat 2.1 "La cerca dels recursos textuals digitals especialitzats sobre els Leònids per a la posterior creació de corpus monlingües comparables" d'aquest mateix capítol (pàgina 167). Així doncs, amb l'ajuda de la funció de cerca del sistema operatiu Windows vam cercar entre els documents hipertextuals descarregats tots aquells arxius amb contingut text que:

- incloguessin la paraula "Leònids", en la llengua corresponent, al cos del recurs i no únicament com a enllaç;
- no tinguessin text en més d'un idioma;
- no fossin missatges electrònics.

Durant aquest procés d'identificació de recursos textuals vàlids vam poder observar evidents transclusions[39] entre llocs web en anglès, generalment provocades per dos motius:

[39] La transclusió, la inclusió d'informació d'un document hipertextual en un altre, es tracta amb més profunditat al final de l'apartat 1.2.2 "Les estructures hipertextuals" del capítol 2, pàgina 76.

- Inclusió en un lloc web d'un recurs textual que forma part d'un altre lloc diferent. És el cas, per exemple, del recurs EN11903 (titulat LEONIDS ON THE MOON), que apareixia íntegre i formalment idèntic al lloc web de la NEW JERSEY ASTRONOMICAL ASSOCIATION (http://www.njaa.org), per la qual cosa el recurs provinent d'aquest segon lloc no va ser inclòs al corpus.
- Repetició de llocs web idèntics amb URL diferents. Aquest fenomen no es pot denominar pròpiament transclusió, però en el corpus final provocaria els mateixos efectes. Aquest és el cas, per exemple, de l'EUROPEAN SOUTHERN OBSERVATORY, que manté dos llocs web idèntics amb diferent adreça URL (http://www.eso.org i http://web1.hq.eso.org).
- Repetició de part de llocs web específics en altres llocs de més genèrics, com pot ser el lloc d'un departament d'una institució repetit en el lloc d'aquesta última. Aquest és el cas, per exemple, de la majoria de departaments o grups de recerca de la NASA, que compten amb un lloc web propi i tornen a oferir part de la informació al lloc web general de la NASA.

Els efectes de la transclusió dins del corpus que hem recollit, així com els de qualsevol altre tipus de repetició innecessària d'informació (tal com es veurà al capítol 6 "L'anàlisi del corpus compilat", pàgina 283) poden provocar alteracions en els resultats de la seva anàlisi, ja que es basa amb la metodologia de la lingüística de corpus que és eminentment quantitativa.

D'altra banda, la cerca de recursos textuals vàlids per al nostre corpus ens va permetre identificar arxius de diferents formats de text: el format HTML, propi dels textos d'Internet; el format DOC, propi del processador de textos MS WORD; i el format TXT, propi d'arxius que contenen text pla sense format. Per evitar que en analitzar els corpus, el programa d'anàlisi prengués les etiquetes HTML com a part del text i, per tant, els resultats de l'anàlisi es veiessin distorsionats, vam convertir cadascun dels recursos acceptats al format TXT. En convertir-los a TXT també els hem denominat a partir de l'identificador que rebran un cop classificats (vegeu l'apartat 2 "La classificació dels recursos textuals digitals especialitzats dedicats als Leònids" del proper capítol,

pàgina 197). La nomenclatura que hem seguit es basa en el codi de la llengua seguit per cinc xifres: les tres primeres relatives al lloc web i les dues últimes al recurs dins d'aquell lloc:

- La llengua: els recursos poden contenir informació en anglès (EN), castellà (ES) o català (CA). En aquest darrer cas, hem diferenciat entre els recursos obtinguts a partir de la cerca de "Leònids" en qualsevol de les seves grafies i dels que ho han estat a partir de la cerca d'algun dels seus hiperònims; en aquest segon cas, la indicació de llengua és CAM (CA + METEORIT).
- El lloc web: els recursos textuals s'han anat cercant per llocs web. Cadascun d'aquests llocs s'identifica amb les primeres tres xifres del nom final de l'arxiu. D'aquesta manera, el lloc EN001 és el primer lloc web en llengua anglesa analitzat, mentre que el lloc ES001 ho és en castellà i el CA001 en català.
- El recurs: les dues darreres xifres del nom de cada recurs corresponen a l'ordre en què ha estat identificat dins del seu lloc web. Així doncs, per exemple, el recurs EN00302 és el segon recurs textual vàlid que hem identificat dins del lloc web 003 en llengua anglesa.

Per tant, el nom dels arxius que formaran part del cadascun dels corpus monolingües s'ha d'interpretar de la manera següent:

EN	001	03	.txt
(llengua)	(lloc web)	(recurs)	(format)

Després de totes aquestes operacions, la distribució des recursos textuals que finalment formaran part dels corpus és aquesta:

Llengües	Recursos textuals
Anglès	922
Castellà	116
Català	77[40]
Total	1115

Taula 3-10. Recursos textuals que formaran part dels corpus monolingües (e. p.)

A causa de la naturalesa diferent dels llocs web dels quals provenen els recursos textuals que finalment formaran part dels corpus monolingües, la quantitat de recursos que aporta cada lloc web varia:

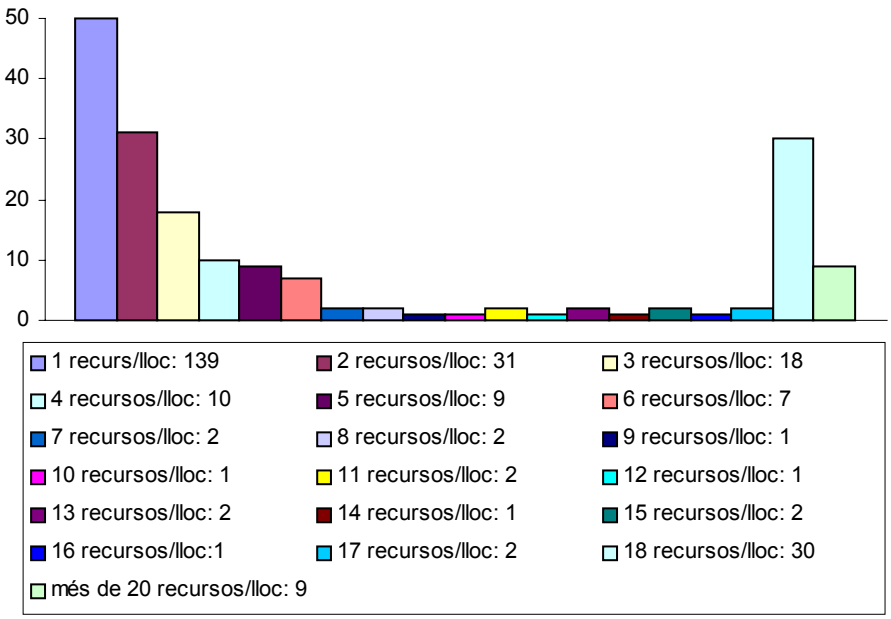

Figura 3-5. Distribució dels recursos textuals segons els llocs web en anglés (e. p.)

[40] Vint-i-un recursos que contenen "Leònids" o variacions de "Leònids" i cinquanta-sis llocs que contenen algun dels hiperònims de Leònids.

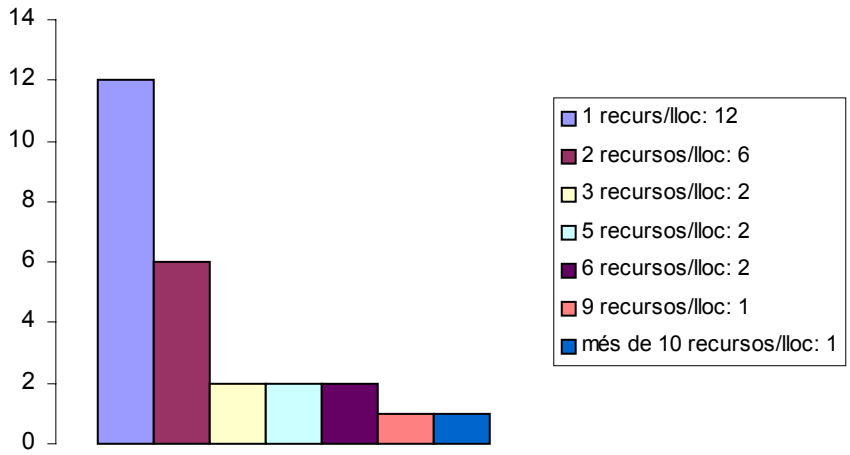

Figura 3-6. Distribució dels recursos textuals segons els llocs web en castellà (e. p.)

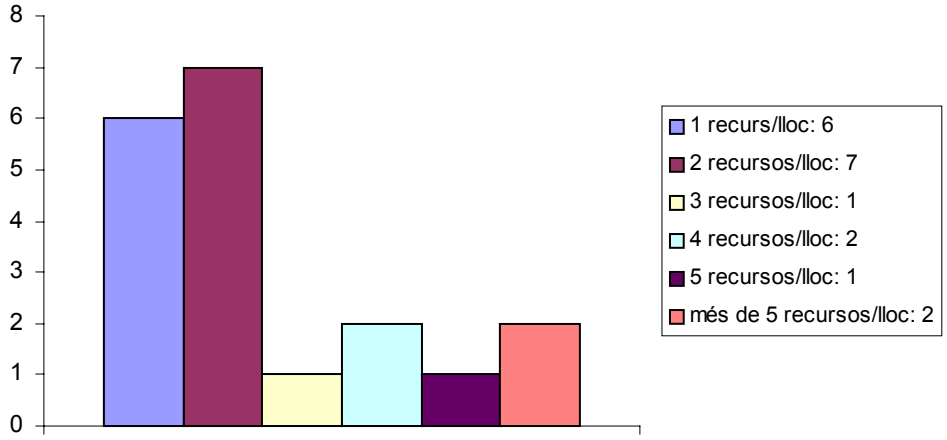

Figura 3-7. Distribució dels recursos textuals segons els llocs web en català (e. p.)

En aquest capítol hem aprofundit en la noció de *World Wide Web*, els criteris que habitualment se segueixen per tal d'identificar-hi recursos de qualitat i les eines que

permeten localitzar-los, és a dir, els diferents tipus de cercadors. Aquest domini de l'entorn del qual es vol extreure textos paral·lels és imprescindible per tal de dur a terme aquesta operació amb èxit. Tanmateix, a més de conèixer l'entorn documental i el seu funcionament, el traductor ha de ser conscient de quines són les seves necessitats informatives. D'aquesta manera, fent convergir les seves necessitats d'informació amb les possibilitats de cerca de recursos textuals digitals, el traductor podrà dissenyar les estratègies de cerca que més s'hi adeqüin.

Per il·lustrar el procés de cerca i obtenció de recursos textuals digitals especialitzats, hem dut a terme una cerca a partir de les paraules clau Leònids i astronomia (en cadascuna de les llengües de treball escollides i tenint en compte possibles variacions ortogràfiques). Els recursos obtinguts integraran el corpus *ad hoc* que utilitzarem com a text paral·lel.